

Debate-aware Learning Makes LLMs Genuine Debaters

Yufei Zhang^{♣†} Kai Xiong^{♡†} Yuxiong Yan[♣] Bibo Cai^{♣*}
Xiao Ding^{♣*} Bing Qin[♣] Ting Liu[♣]

[♣]Research Center for Social Computing and Interactive Robotics
Harbin Institute of Technology, Harbin, China

[♡]Zhongguancun Laboratory, Beijing, China

{yfzang, yxian, bbcai, xding, qinb, tliu}@ir.hit.edu.cn
xiongk@zgclab.edu.cn

Abstract

Multi-agent debate systems are used to competitively discuss and resolve complex tasks. However, existing approaches often exhibit pseudo-debate behavior: agents either ignore opponents' arguments or compromise too readily without critical evaluation. To address this, we propose a three-stage debate framework grounded in abstract argumentation theory. First, Debate Structuring uses principled prompts to elicit well-formed debate interactions. Second, Debate Imitation filters high-quality debates to fine-tune LLMs on proper argumentative conduct. Third, Debate Internalization employs a debate-aware reinforcement learning objective to instill adversarial reasoning patterns. To better quantify debate capability, we introduce two novel metrics: Rebuttal Engagement Score (RES), assessing how thoroughly models engage with opposing arguments, and Belief Governance Score (BGS), evaluating whether models appropriately retain or revise beliefs based on argument strength. Extensive evaluations across six benchmarks show that our framework not only improves task performance but also cultivates more authentic and robust debate behavior in LLMs. Our work contributes a theory-inspired training framework and evaluation suite for advancing multi-agent debate systems.

1 Introduction

Large Language Models (LLMs) have demonstrated remarkable problem-solving capabilities (Tang et al., 2024; Xiong et al., 2024) and have been applied across a wide range of domains (Zhao et al., 2024; Shool et al., 2025). To further solve complex tasks, multi-agent debate systems (Chan et al., 2024; Becker et al., 2025) are proposed to discuss and solve problems in a competitive way, expecting LLM-based agents can achieve a uni-

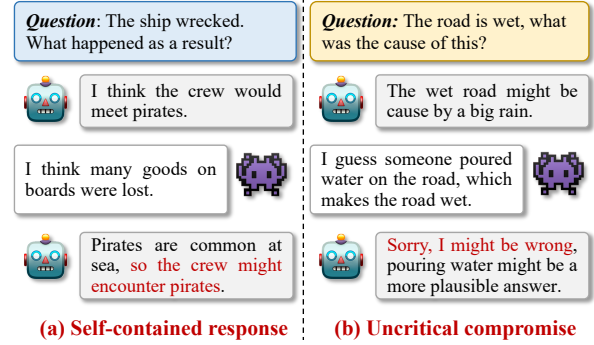


Figure 1: (a) Self-contained response and (b) uncritical compromise in multi-agent systems.

fied objective through alternative presentation of evidence and rebuttal.

Existing multi-agent debate methods (Xiong et al., 2023; Liang et al., 2024) mainly use predefined stance to prompt LLMs to debate or generate debate data for fine-tuning. Although they have made some progress, LLMs in multi-agent debate systems often exhibit a pseudo-debate behavior, where participants assume opposing roles within a debate-like structure without engaging in a genuine adversarial process. As shown in Figure 1, this phenomenon manifests in two aspects. On the one hand, LLMs fail to fully take other agents' opinions into consideration, frequently responding in a self-contained manner that prioritizes their own prior reasoning over the newly introduced arguments. On the other hand, LLMs are prone to readily conceding to others' opinions without sufficient scrutiny, exhibiting a tendency to compromise or conform even when the opposing arguments are weak or flawed.

The pseudo-debate behavior is mainly due to the training strategies of LLMs. They encourage LLMs to become strong independent problem solvers (Liu et al., 2024; Guo et al., 2025), leaving little room for learning how to properly engage with and critically evaluate others' arguments. Further-

[†]Equal Contribution

^{*}Corresponding Authors

more, some training paradigms such as RLHF (Bai et al., 2022) and DPO (Rafailov et al., 2023) would implicitly bias LLMs toward being overly agreeable conversational partners. Under such settings, LLMs are prone to uncritically conceding to opposing views without sufficient scrutiny.

To solve the above problems, we follow the Abstract Argumentation Frameworks (Dung, 1995) and design a three-stage debate framework in this paper: (1) Debate Structuring utilizes a carefully devised prompt grounded in debate principles to elicit well-formed, theory-guided debate interactions between LLMs; (2) Debate Imitation filters high-quality debate trajectories generated in the first stage and employs supervised fine-tuning to teach LLMs to reproduce proper debate behaviors; (3) Debate Internalization further introduces a debate-aware reinforcement learning algorithm to internalize adversarial reasoning patterns, enabling LLMs to engage in genuine argumentative exchanges rather than superficial role-playing.

Furthermore, to better measure the debate ability of LLMs, we design two metrics: (1) a Rebuttal Engagement Score (RES) to evaluate whether the model truly engages with the opponent’s arguments, (2) a Belief Governance Score (BGS) to measure whether the model holds its ground when correct and revises its stance when mistaken.

We conduct extensive experiments across six reasoning and language understanding benchmarks. The results consistently demonstrate the superiority of our proposed three-stage debate framework. Moreover, performance on our two newly introduced metrics further indicates that our method enables LLMs to engage in more human-like debate behaviors, moving beyond superficial role-playing toward genuine adversarial reasoning. The code will be made publicly available at [GitHub](#).

Our contributions are summarized as follows:

- We adopt debate theory and design a three-stage debate framework to address the pseudo-debate issues in multi-agent debate systems.
- We propose two new metrics to quantify the debate capabilities of LLMs.
- Our proposed framework could achieve better performance and stronger debate skills.

2 Method

In this section, we present the proposed three-stage debate framework, as shown in Figure 2.

The framework first structures open-ended interaction into commitment-centered attack-defense moves, then filters high-quality trajectories for supervised imitation learning, and finally applies debate-oriented reinforcement learning to internalize when to maintain, revise, or challenge beliefs.

2.1 Problem Formulation

Given a reasoning question x with candidate options O , the goal of debate-based reasoning is to obtain the final answer y^* through multi-agent interaction. A debate process is represented as a trajectory:

$$\tau = \{u_1^0, u_2^0, u_1^1, u_2^1, \dots, u_1^T, u_2^T\},$$

where u_i^t denotes the utterance of agent i at turn t , including its commitments, argumentative moves, and current answer. Different from conventional debate methods that only optimize final answer accuracy, we aim to train agents that can appropriately maintain, revise, or challenge beliefs during the interaction.

2.2 Debate Structuring

To address pseudo-debate behavior, we first introduce Debate Structuring. Grounded in Abstract Argumentation Theory, this stage organizes the interaction around claims, supporting grounds, attacks, and defenses, requiring each response to expose its current beliefs and explicitly engage with the opponent’s arguments.

For each non-initial turn, the model receives the task input, the answer options, and the full debate history. Its output is a structured debate move:

$$u_i^t = (\mathcal{D}_i^t, \mathcal{A}_i^t, \mathcal{C}_i^t, y_i^t), \quad t \geq 1. \quad (1)$$

Here, $\mathcal{D}_i^t$ denotes the defense part, $\mathcal{A}_i^t$ denotes the attack part, $\mathcal{C}_i^t$ denotes the updated commitment set, and y_i^t denotes the current answer. The defense part asks the model to respond to previous challenges, explaining whether the attacked commitment should be kept, modified, or retracted. The attack part asks the model to challenge the opponent’s commitments or reasoning, such as factual errors, invalid reasoning links, unsupported rules, counterexamples, or missing conditions. The updated commitment set records the beliefs the model still accepts after considering the debate, and the final answer should be consistent with these updated commitments.

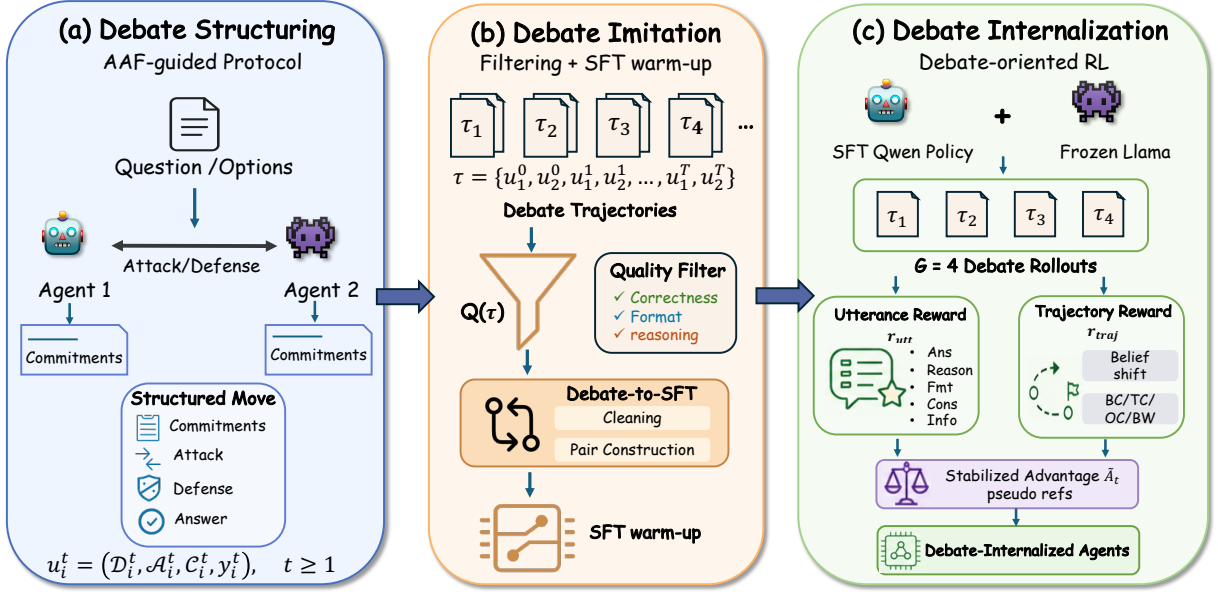


Figure 2: Overview of the three-stage-debate framework. (a) Debate Structuring converts open-ended dialogue into structured argumentative moves grounded in commitments, attacks, and defenses. (b) Debate Imitation filters high-quality debate trajectories and uses them for supervised fine-tuning as a warm-up for the debate protocol. (c) Debate Internalization applies debate-oriented reinforcement learning to optimize argumentative actions and belief transitions, enabling the model to learn when to maintain, revise, or challenge beliefs.

The initial turn is treated as a special case because no complete debate history is available. The first agent independently produces initial commitments and an answer, while the second agent first forms its own answer and then may rebut the first agent’s commitments. After initialization, agents take turns producing structured moves based on the accumulated history, forming a debate trajectory:

$$\tau = \{u_1^0, u_2^0, u_1^1, u_2^1, \dots, u_1^T, u_2^T\}. \quad (2)$$

The debate stops when the agents reach answer agreement or the maximum number of turns is reached.

This protocol keeps the debate close to the theoretical requirements of argumentation: each answer is supported by commitments, each disagreement targets a specific claim, and each challenged belief must be defended or revised. The complete prompt templates are provided in [Appendix A](#).

2.3 Debate Imitation

To transfer the structured behaviors elicited in Debate Structuring into the model itself, we introduce Debate Imitation as the second stage. It filters trajectories that are both outcome-correct and argumentatively well-formed, and uses them for supervised fine-tuning as a warm-up before reinforcement learning.

For each trajectory τ defined in Section 2.2, we compute a quality score:

$$Q(\tau) = Q_{\text{ans}}(\tau) + Q_{\text{fmt}}(\tau) + Q_{\text{arg}}(\tau) + Q_{\text{contrib}}(\tau). \quad (3)$$

Here, Q_{ans} measures final-answer correctness, Q_{fmt} checks whether the required structured fields are present, Q_{arg} verifies whether attacks and defenses are grounded in the corresponding commitments, and Q_{contrib} rewards useful debate behaviors such as valid rebuttal, belief revision, belief maintenance, and acceptance of stronger arguments. We retain trajectories whose overall quality satisfies the filtering criteria. Detailed filtering rules and turn-level SFT construction are provided in [Appendix B](#).

From the selected trajectories, we construct turn-level imitation examples. For each agent, the input state s_t consists of the original question, options, and the debate history before the current turn, while the target output u_t is the agent’s structured response. Each model is trained only on its own utterances, with the other agent’s responses treated as dialogue context. The supervised objective is:

$$\mathcal{L}_{\text{SFT}} = - \sum_{(s_t, u_t) \in \mathcal{D}_{\text{sft}}} \log \pi_{\theta}(u_t | s_t). \quad (4)$$

This stage teaches the model to follow the structured debate protocol, but it does not directly opti-

mize when to maintain, revise, or reject beliefs. We therefore introduce debate-oriented reinforcement learning in the next stage.

2.4 Debate Internalization

Although imitation learning provides well-formed debate examples, it does not directly optimize when to maintain, revise, or challenge beliefs during interaction. We therefore introduce debate-oriented reinforcement learning. The SFT-trained Qwen model is optimized as the trainable policy, while the SFT-trained Llama model is kept frozen as the opponent. For each training instance x , we sample $G = 4$ complete debate rollouts:

$$\mathcal{G}(x) = \{\tau_g\}_{g=1}^G. \quad (5)$$

Each rollout contains the full interaction between the trainable agent p and the opponent o . Rewards are computed over the full debate process, but policy gradients are applied only to the trainable agent’s utterances.

For each utterance u_i^t , we first compute an utterance-level reward that evaluates the local quality of a single debate move:

$$r_{\text{utt}}(u_i^t) = w_1 r_{\text{ans}} + w_2 r_{\text{reason}} + w_3 r_{\text{fmt}} + w_4 r_{\text{cons}} + w_5 r_{\text{info}}. \quad (6)$$

This reward combines answer correctness with process-level debate quality. Specifically, r_{reason} captures belief maintenance, rebuttal, compromise, or correction; r_{fmt} checks structural compliance; r_{cons} verifies attack-defense grounding; and r_{info} encourages useful non-redundant information.

In addition to local utterance quality, we define a trajectory-level reward to evaluate whether the whole debate leads to a desirable belief transition. Let y^* denote the ground-truth answer, and let $y_{p,g}^0$ and $y_{o,g}^0$ denote the initial answers of the trainable agent and the opponent in rollout τ_g . We categorize each rollout by the initial correctness pattern:

$$c_g = \begin{cases} \text{BC}, & y_{p,g}^0 = y^*, y_{o,g}^0 = y^*, \\ \text{TC}, & y_{p,g}^0 = y^*, y_{o,g}^0 \neq y^*, \\ \text{OC}, & y_{p,g}^0 \neq y^*, y_{o,g}^0 = y^*, \\ \text{BW}, & y_{p,g}^0 \neq y^*, y_{o,g}^0 \neq y^*. \end{cases} \quad (7)$$

These categories correspond to different desired belief transitions. Both Correct (BC) emphasizes stability, since both agents start correctly. Trainable Correct (TC) requires the trainable agent to resist invalid compromise, while Opponent Correct (OC)

requires it to revise its belief by accepting the correct opposing argument. Both Wrong (BW) evaluates whether the debate can recover from initially wrong answers. We therefore define the trajectory reward as a category-aware function:

$$r_{\text{traj}}(\tau_g) = F(c_g, y_g^T, \text{Agree}_g), \quad (8)$$

where y_g^T is the final answer of the rollout, and Agree_g indicates whether the two agents reach agreement. This reward is used only for rollout-level advantage estimation, rather than being directly assigned to any single utterance.

To combine local and global signals, we compute a train-side mean reward for each rollout by averaging the utterance-level rewards over the trainable agent’s turns:

$$\bar{r}_{\text{train}}(\tau_g) = \frac{1}{|\mathcal{T}_p|} \sum_{t \in \mathcal{T}_p} r_{\text{utt}}(u_{p,g}^t), \quad (9)$$

where $\mathcal{T}_p$ denotes the set of trainable-agent turns. The rollout score is then defined as:

$$s(\tau_g) = 0.7 \times r_{\text{traj}}(\tau_g) + 0.3 \times \bar{r}_{\text{train}}(\tau_g). \quad (10)$$

This design allows the trajectory-level reward to determine whether the overall debate follows a desirable belief transition, while the train-side mean reward reflects the local quality of the trainable agent’s own utterances. The concrete utterance-level and trajectory-level reward rules are provided in [Appendix C](#).

Standard group-relative normalization can be unstable when all rollouts share the same correctness outcome. We therefore stabilize the advantage computation by adding a pseudo reference only to the group statistics: mixed groups use the original rollout scores, all-correct groups add a pseudo-negative score s^- , and all-wrong groups add a pseudo-positive score s^+ . Let $\mathcal{S}'$ denote the resulting score set. The rollout advantage is computed as:

$$A_g = \frac{s(\tau_g) - \mu_{\mathcal{S}'}}{\sigma_{\mathcal{S}'} + \epsilon}. \quad (11)$$

The pseudo references are used only to compute $\mu_{\mathcal{S}'}$ and $\sigma_{\mathcal{S}'}$, and are not treated as real training samples.

Finally, we assign credit to each trainable-agent utterance. Let

$$c_{g,t} = r_{\text{utt}}(u_{p,g}^t) \quad (12)$$

denote the local utterance credit. The effective advantage for utterance $u_{p,g}^t$ is:

$$\tilde{A}_{g,t} = A_g \cdot c_{g,t}. \quad (13)$$

Thus, the rollout advantage determines whether the whole trajectory should be encouraged or discouraged, while the utterance-level credit controls how strongly each trainable-agent turn contributes to the update.

We optimize the trainable policy using a clipped policy objective with KL regularization against the SFT reference policy. The loss is applied only to the trainable agent’s utterance tokens, while the frozen opponent is used only to generate debate context and rewards. In this way, the model is optimized not only for answer correctness, but also for maintaining, revising, and defending beliefs through explicit argumentative interaction.

3 Experiments

3.1 Evaluation Metrics

Task accuracy only measures whether the final answer is correct, but it does not reveal whether the model genuinely engages in debate. We therefore introduce two process-level metrics for the trainable agent: Belief Governance Score (BGS), which evaluates whether the model updates its beliefs properly, and Rebuttal Engagement Score (RES), which evaluates whether the model substantively responds to the opponent.

3.1.1 Belief Governance Score

BGS measures whether the model can maintain correct beliefs, revise incorrect ones, and avoid unstable answer changes. For each sample, we extract the answer sequence from the debate context and compare each answer with the ground-truth answer y^* . According to the correctness of the first two answers, we categorize each debate into four initial states: both agents correct, only the trainable agent correct, only the opponent correct, and both agents wrong. Based on these states, we compute state transition reliability:

$$\begin{aligned} \text{STR} = & 0.6 \cdot \text{Keep}_{BC} + 0.3 \cdot \text{PairChoice}_{TO} \\ & + 0.1 \cdot \text{Recover}_{BW}. \end{aligned} \quad (14)$$

Here, Keep_{BC} is the final accuracy when both agents are initially correct, and Recover_{BW} is the final accuracy when both agents are initially

wrong. PairChoice_{TO} measures whether the trainable agent selects the correct side when exactly one agent is initially correct; it is computed as the sample-count-weighted average of Resist_{TC} and Learn_{OC} , where Resist_{TC} measures final accuracy when only the trainable agent is initially correct, and Learn_{OC} measures final accuracy when only the opponent is initially correct.

We also compute selective update reliability, which measures whether answer changes are valid and stable:

$$\text{SUR} = \frac{\text{VSR} + \text{TSR} + \text{OSR}}{3}. \quad (15)$$

VSR is the valid shift rate and is computed only over transitions where the answer changes: wrong-to-correct receives 1, correct-to-wrong receives 0, and wrong-to-wrong receives 0.5. TSR is the terminal stability rate: it is 1 if the last two answers are identical, 0.5 if the final answer changes only once, and 0 if the final stage remains unstable. OSR is the oscillation suppression rate, defined as $1 - \text{OR}$, where OR maps zero or one answer change to 0, two changes to 0.5, and three or more changes to 1.

The final BGS is defined as:

$$\text{BGS} = 0.50 \cdot \text{STR} + 0.50 \cdot \text{SUR}. \quad (16)$$

3.1.2 Rebuttal Engagement Score

RES measures whether the model substantively engages with the opponent’s arguments. For each trainable-agent response turn, we extract attack and defense units and compare them with the previous turn’s commitments and argument units. We first compute:

$$\text{RES}_{core} = \frac{\text{TH} + \text{UR} + \text{FC}}{3}. \quad (17)$$

Here, TH measures whether the current attack or defense targets the opponent’s previous commitments or core arguments, UR measures whether the model gives a substantive response to the targeted argument, and FC measures whether the debate remains focused on the same disputed point. These scores are computed by rule-based extraction and semantic similarity matching, and mapped to $\{0, 0.5, 1\}$.

To discourage overly short or empty rebuttals, we also include an argument-unit quantity score ADQ, defined as the clipped ratio between the number of extracted attack/defense units and the expected number of units. The final RES is:

$$\text{RES} = (1 - \eta)\text{RES}_{core} + \eta\text{ADQ}. \quad (18)$$

In our implementation, $\eta = 0.15$. All RES components are averaged over valid trainable-agent response turns. The calculation does not rely on an LLM judge.

3.2 Evaluation Benchmarks

To comprehensively assess reasoning and language understanding capabilities of all methods across diverse domains, we select 6 widely used evaluation benchmarks for evaluation: (1) an abductive reasoning benchmark α NLI (Bhagavatula et al., 2020); (2) a language understanding benchmark MMLU (Hendrycks et al., 2021); (3) a natural-language-form logical reasoning benchmark LogiQA2 (Liu et al., 2023); (4) a causal reasoning benchmark e-CARE (Du et al., 2022); (5) a complex reasoning benchmark BIG-Bench (Srivastava et al., 2023); (6) an analytical reasoning benchmark AR-LSAT (Zhong et al., 2021). Together, these benchmarks provide a representative and challenging set of tests for inference, knowledge, logic, causality, and structured problem solving.

3.3 Baselines

We adopt a series of baselines for comparisons: (1) Single LLM utilizes base LLMs to answer questions with CoT, we use Qwen-3 (Yang et al., 2025) and LLaMA-3 (Grattafiori et al., 2024) for experiments; (2) Majority Voting (Choi et al., 2026) follows majority voting across multiple reasoning runs to obtain the final answers; (3) Self-consistency (Wang et al., 2023) adopt majority voting across multiple rollouts to obtain the final answers; (4) FORD (Xiong et al., 2023) adopts a formal debate framework to solve problems with a carefully designed prompt; (5) MAD (Liang et al., 2024) employs a debate prompt to answer questions, similar to FORD; (6) GRPO (Guo et al., 2025) uses rule-based rewards for RL training. All baselines are reimplemented and evaluated under the same experimental setting. We use the same test instances, backbone models, decoding strategies, and evaluation protocols for fair comparison.

3.4 Implementation Details

3.4.1 Debate Structuring

We use Qwen3-4B-Instruct-2507 and Meta-Llama-3-8B-Instruct as the two debating agents. For each task instance, the agents conduct a structured debate following the protocol in Section 2.2. The maximum number of utterances is set to 6, and the debate stops early once the two agents reach

answer agreement. During debate generation, we use temperature 0, top- p 0.9, and a maximum generation length of 4096 tokens. The same decoding setting is used during evaluation.

3.4.2 Debate Imitation

For Debate Imitation, we fine-tune both Qwen and Llama on the filtered debate trajectories. Each model is trained only on its own utterances, while the other agent’s utterances are used as dialogue context. The input consists of the question, answer options, and previous debate history, and the target is the current structured response. We train both models with the same SFT setting: learning rate 1×10^{-5} , batch size 4, 3 epochs, warmup steps 50, and maximum sequence length 5120.

3.4.3 Debate Internalization

For Debate Internalization, we initialize the trainable policy from the SFT-trained Qwen checkpoint and use the SFT-trained Llama as a frozen opponent. For each training instance, we sample 4 complete debate rollouts with maximum 6 utterances, temperature 0.8, top- p 0.95, and maximum 800 new tokens per utterance. The rollout score combines the dialogue-level reward and the train-side utterance reward with weights 0.7 and 0.3. We update only Qwen utterance tokens using clipped policy optimization with KL regularization. The RL training uses learning rate 5×10^{-6} , KL coefficient 0.02, clipping threshold 0.2, update epochs 2, minibatch size 4 utterances, and gradient accumulation 2. We use checkpoint-1200 for final evaluation. All training is conducted on A100-80GB GPUs. The total computational budget is approximately 450 A100-80GB GPU hours. Detailed computational budget, package settings, and implementation parameters are provided in Appendix D.3 and Appendix D.4.

4 Results

We have evaluate reasoning and debate capabilities of all methods, results are shown in Table 1. Unless otherwise stated, we report single-run results due to the computational cost of debate-oriented reinforcement learning. We can have the following observations:

4.1 Reasoning Capability: Accuracy

As shown in Table 1 (a):

(1) Our proposed three-stage framework consistently improves the overall performance across six benchmarks. In particular, Debate Internalization

Method	α NLI	MMLU	LogiQA2	e-CARE	BIG-bench	AR-LSAT	Average
(a) Accuracy $\uparrow$							
Qwen3 (Yang et al., 2025)	78.4	77.2	73.2	82.8	91.6	46.4	74.9
LLaMA-3 (Grattafiori et al., 2024)	72.4	71.2	56.4	77.6	35.6	22.8	56.0
Majority Voting (Choi et al., 2026)	79.2	80.4	71.6	83.2	86.8	54.8	76.1
Self-consistency (Wang et al., 2023)	78.0	79.6	74.0	82.0	91.6	56.0	76.9
FORD (Xiong et al., 2023)	76.0	76.8	67.2	80.0	90.4	61.6	75.3
MAD (Liang et al., 2024)	78.4	80.0	64.8	80.0	93.6	58.8	75.8
GRPO (Guo et al., 2025)	78.4	78.4	74.0	79.6	92.4	64.8	77.9
Debate Structuring (Ours)	78.4	77.6	70.8	79.2	93.6	67.6	77.8
Debate Imitation (Ours)	78.4	77.6	70.8	79.6	93.2	64.4	77.3
Debate Internalization (Ours)	75.6	80.4	74.4	81.2	94.4	68.0	79.0
(b) Rebuttal Engagement Score (RES) $\uparrow$							
FORD (Xiong et al., 2023)	88.8	65.6	43.9	63.3	47.2	48.7	59.6
MAD (Liang et al., 2024)	88.5	66.6	45.8	66.3	49.3	50.8	61.2
GRPO (Guo et al., 2025)	78.9	79.7	77.8	78.5	73.4	62.8	75.2
Debate Structuring (Ours)	72.3	65.9	54.5	72.5	61.9	52.4	63.2
Debate Imitation (Ours)	77.7	78.6	76.6	77.0	72.7	70.5	75.5
Debate Internalization (Ours)	82.0	81.8	81.1	80.6	77.6	70.1	78.9
(c) Belief Governance Score (BGS) $\uparrow$							
FORD (Xiong et al., 2023)	76.6	79.0	76.0	76.4	79.6	77.6	77.5
MAD (Liang et al., 2024)	77.0	80.7	73.9	77	82.4	77.6	78.1
GRPO (Guo et al., 2025)	83.5	79.6	77.9	79.1	90.2	73.6	80.6
Debate Structuring (Ours)	78.8	79.7	77.1	79.8	88.9	76.5	80.1
Debate Imitation (Ours)	80.4	80.0	75.8	76.3	89.4	75.4	79.6
Debate Internalization (Ours)	88.2	80.5	78.9	73.9	86.0	80.1	81.3

Table 1: Overall evaluation results across 6 benchmarks. The improvements of Debate Internalization over the strongest baseline are statistically significant under paired t-test ($p < 0.05$).

achieves the best average accuracy, outperforming all baseline methods.

(2) Existing debate methods such as FORD and MAD yield unstable improvements, suggesting that genuine debate requires more careful guidance.

(3) Debate Structuring already outperforms most baselines, this is because that theory-guided prompts can substantially improve debate quality.

(4) Our framework achieves particularly large improvements on challenging reasoning benchmarks such as AR-LSAT. This suggests that proper debate behaviors are especially beneficial for complex logical reasoning.

4.2 Debate Capability: Rebuttal Engagement

As shown in Table 1 (b):

(1) Debate Internalization achieves the highest RES, indicating that debate-aware RL helps models better engage with opponents’ arguments.

(2) Debate Imitation significantly improves RES over Debate Structuring, showing that high-quality debate trajectories effectively teach rebuttal behaviors.

(3) FORD and MAD achieve low RES, suggesting that naive role-playing cannot guarantee deep argumentative interaction.

(4) Debate Structuring already improves RES on datasets such as e-CARE and BIG-bench, demon-

strating that theory-guided prompts can encourage more opponent-aware responses.

(5) The consistent RES improvements across all benchmarks indicate that our framework effectively alleviates the self-contained response problem.

4.3 Debate Capability: Belief Governance

As shown in Table 1 (c):

(1) Debate Internalization achieves the best BGS, indicating that debate-aware RL improves the model’s ability to maintain or revise beliefs appropriately.

(2) Debate Structuring already surpasses FORD and MAD on most datasets, suggesting that theory-guided debate prompts can reduce uncritical compromise.

(3) Debate Imitation does not consistently improve BGS, showing that supervised fine-tuning alone is insufficient for learning robust belief governance behaviors.

(4) Our method achieves particularly strong gains on α NLI and AR-LSAT, because these tasks require careful evaluation of conflicting arguments.

(5) The consistent improvement of BGS demonstrates that our framework enables models to critically evaluate opponents’ arguments instead of blindly conceding.

Method	Accuracy
MAD	75.8
Debate Imitation + MAD	75.9
Debate Internalization + MAD	76.7

Table 2: Transfer capabilities on different prompt.

Method	Accuracy
Qwen3	76.8
Majority Voting	75.3
Self-consistency	81.3
MAD	81.8
Debate Internalization (Ours)	84.9

Table 3: OOD generalization results.

5 Analysis

5.1 Debate Imitation Transfer Analysis

We conduct this analysis to investigate the generalization ability of Debate Imitation to other debate prompts. Specifically, we adopt the prompt of MAD (Liang et al., 2024) to debate based on the LLMs after Debate Imitation. The results are shown in Table 2. LLMs after Debate Imitation can properly applied to other debate prompts, indicating that the debate capability learned from our structured debate data is transferable rather than overfitted to a specific prompting protocol.

5.2 Debate Internalization Transfer Analysis

Specifically, we further apply the LLMs after Debate Internalization to the debate framework of MAD (Liang et al., 2024). As shown in Table 2, Debate Internalization achieves a larger improvement compared with Debate Imitation, demonstrating that debate-aware RL further strengthens the model’s intrinsic debate capability. These results suggest that the debate behaviors learned in our framework can generalize across different debate prompting protocols, rather than being tightly coupled with our proposed debate structure.

5.3 Out-of-domain Generalization

To further investigate the generalization performance of our proposed method, we choose three out-of-domain math benchmarks for OOD evaluation: (1) AMC 12^{*}; (2) CompMath (Raimondi et al., 2026); (3) TAL-ACQ5k[†]. The averaged performance across all benchmark are shown in

^{*}<https://huggingface.co/datasets/edev2000/amc12-full>

[†]<https://github.com/math-eval/TAL-SCQ5K>

Setting	Metric	Agreement
DI(ours) vs. GRPO	–	67.50% (81/120)
Same-question	BGS	82.50% (99/120)
Same-question	RES	80.83% (97/120)
Cross-question	BGS	76.67% (46/60)
Cross-question	RES	73.33% (44/60)

Table 4: Human validation of debate quality metrics and method preference.

Table 3, our method achieves the best OOD performance. The results indicate that Debate Internalization enables LLMs to acquire more robust and transferable reasoning behaviors, leading to stronger generalization under distribution shifts.

5.4 Human Validation

To validate whether RES and BGS accurately reflect human judgments of debate quality, we conduct a blinded human evaluation. We denote our full method, Debate Internalization, as DI for brevity. We randomly sample 20 questions from each of the six benchmarks, resulting in 120 questions in total, and collect the corresponding debate trajectories produced by DI and GRPO. Three annotators with backgrounds in LLM reasoning and debate independently compare the anonymized debate results.

For each comparison, annotators evaluate three aspects: overall debate quality, rebuttal persuasiveness, and the appropriateness of belief updates. The first dimension is used for direct preference comparison between DI and GRPO, while the latter two are used to validate RES and BGS, respectively. We additionally construct 60 cross-question pairs within Debate Internalization to evaluate metric-human agreement beyond direct method comparisons. Detailed annotation instructions, recruitment and compensation information, and consent procedures are provided in Appendix D.5.

As shown in Table 4, both proposed metrics show strong agreement with human judgments. BGS and RES achieve 82.50% and 80.83% agreement in same-question comparisons, and maintain 76.67% and 73.33% agreement in cross-question rankings. These indicate that the proposed metrics are not merely capturing differences between methods but also reflect human-perceived debate quality. Moreover, human evaluators prefer Debate Internalization over GRPO in 67.50% of direct comparisons, providing additional evidence that our method improves argumentative interaction.

6 Related Works

6.1 Multi-agent Debate for Reasoning

To address complex tasks, multi-agent debate is proposed to discuss in a competitive way.

Xiong et al. (2023) adopted formal debate framework to employ dual debate and roundtable debate in LLMs, achieving performance gain in various commonsense reasoning benchmarks. Du et al. (2024) presented a multi-agent debate framework, in which multiple agents independently generate proposals and collaboratively engage in deliberation on their respective reasoning processes to reach a consensus on a unified response. Liang et al. (2024) further proposed to encourage divergent thinking in LLMs with debate workflow. Samanta et al. (2025) utilized multi-agent debate to iterative improve LLMs. Wynn et al. (2025) analyzed the failure modes in multi-agent debate. Zhang and Xiong (2025) adopted multi-agent debate for improved fine-grained reasoning in math. Wu et al. (2025) discussed debate capability of LLMs in logical reasoning scenario. Zhang et al. (2026) made LLMs play different roles for better debate.

We propose a three-stage framework to address the pseudo-debate issues in existing multi-agent debate frameworks.

6.2 Multi-agent Debate for Application

Since multi-agent debate can exchange opinions among different expert agents, it is employed to solve complex tasks in real applications.

Jeptoo and Sun (2024) and Liu et al. (2025) utilized multi-agent debate for fake news detection. Zhou and Chen (2025) and Du et al. (2025) focused on education area and enhanced it with multi-agent debate. Smit et al. (2023) and Lu et al. (2026) respectively evaluated and improved multi-agent debate frameworks for medical question answering, while Zhao et al. (2025) proposed a layered debating multi-agent system for similar disease diagnosis. Furthermore, some works conduct software engineering with multi-agent debate frameworks (Li et al., 2025; Lin et al., 2025).

We focus on the fundamental debate capabilities of LLMs, which can be used to improve various applications with multi-agent debate.

7 Conclusion

In this paper, we investigate the limitations of existing multi-agent debate systems and identify pseudo-debate behaviors as a major obstacle to effective

collaborative reasoning. To address this issue, we propose a three-stage debate framework consisting of Debate Structuring, Debate Imitation, and Debate Internalization, which progressively guides LLMs from external debate prompting to intrinsic adversarial reasoning. Furthermore, we introduce two new metrics, RES and BGS, to quantitatively evaluate debate quality beyond final task accuracy. Experimental results across multiple benchmarks demonstrate that our framework not only improves reasoning performance, but also enables LLMs to engage in more critical, responsive, and belief-consistent debate behaviors. We hope our work can provide a foundation for developing more authentic and theoretically grounded multi-agent debate systems in future research.

8 Acknowledgements

The research in this article is supported by the New Generation Artificial Intelligence of China (2024YFE0203700), National Natural Science Foundation of China under Grants 62576124, National Cyber Security-National Science and Technology Major Project (2025ZD1503800).

Limitations

Although our framework improves both debate quality and task performance, several limitations remain. First, our method mainly focuses on two-agent debate settings, while more complex multi-party interactions are not fully explored. Second, the proposed RES and BGS metrics primarily evaluate interaction behaviors at the response level and may not completely capture deeper argumentative structures. In addition, the reinforcement learning stage introduces additional training cost compared with prompt-based debate methods. Finally, although our framework alleviates self-contained responses and uncritical compromise, models may still exhibit suboptimal belief updates when facing highly ambiguous or misleading arguments.

References

- Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, and 1 others. 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*.
- Jonas Becker, Lars Benedikt Kaesberg, Niklas Bauer, Jan Philip Wahle, Terry Ruas, and Bela Gipp. 2025.

- Mallm: Multi-agent large language models framework. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 418–439.
- Chandra Bhagavatula, Ronan Le Bras, Chaitanya Malaviya, Keisuke Sakaguchi, Ari Holtzman, Hannah Rashkin, Doug Downey, Wen-tau Yih, and Yejin Choi. 2020. Abductive commonsense reasoning. In *International Conference on Learning Representations*.
- Chi-Min Chan, Weize Chen, Yusheng Su, Jianxuan Yu, Wei Xue, Shanghang Zhang, Jie Fu, and Zhiyuan Liu. 2024. Chateval: Towards better llm-based evaluators through multi-agent debate. In *International conference on learning representations*, volume 2024, pages 9079–9093.
- Hyeong Kyu Choi, Jerry Zhu, and Sharon Li. 2026. Debate or vote: Which yields better decisions in multi-agent large language models? *Advances in Neural Information Processing Systems*, 38:101732–101764.
- Jing Du, Guangtao Xu, Wenhao Liu, Dibin Zhou, and Fuchang Liu. 2025. Enhancing online learning through multi-agent debates for cs university students. *Applied Sciences*, 15(11):5877.
- Li Du, Xiao Ding, Kai Xiong, Ting Liu, and Bing Qin. 2022. e-care: a new dataset for exploring explainable causal reasoning. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 432–446.
- Yilun Du, Shuang Li, Antonio Torralba, Joshua B Tenenbaum, and Igor Mordatch. 2024. Improving factuality and reasoning in language models through multiagent debate. In *Proceedings of the 41st International Conference on Machine Learning*, pages 11733–11763.
- Phan Minh Dung. 1995. On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. *Artificial intelligence*, 77(2):321–357.
- Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.
- Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shiron Ma, Xiao Bi, and 1 others. 2025. Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. *Nature*, 645(8081):633–638.
- Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. In *International Conference on Learning Representations*.
- Korir Nancy Jeptoo and Chengjie Sun. 2024. Enhancing fake news detection with large language models through multi-agent debates. In *CCF International Conference on Natural Language Processing and Chinese Computing*, pages 474–486. Springer.
- Han Li, Yuling Shi, Shaoxin Lin, Xiaodong Gu, Heng Lian, Xin Wang, Yantao Jia, Tao Huang, and Qianxiang Wang. 2025. Swe-debate: Competitive multi-agent debate for software issue resolution. *arXiv preprint arXiv:2507.23348*.
- Tian Liang, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, Shuming Shi, and Zhaopeng Tu. 2024. Encouraging divergent thinking in large language models through multi-agent debate. In *Proceedings of the 2024 conference on empirical methods in natural language processing*, pages 17889–17904.
- Kaicun Lin, Tao Zou, and Huaqiang Yuan. 2025. Debate, verify, and debug: A multi-agent planning framework for reliable code generation. In *2025 IEEE 8th International Conference on Computer and Communication Engineering Technology (CCET)*, pages 1–7. IEEE.
- Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, and 1 others. 2024. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.
- Hanmeng Liu, Jian Liu, Leyang Cui, Zhiyang Teng, Nan Duan, Ming Zhou, and Yue Zhang. 2023. Logiqa 2.0—an improved dataset for logical reasoning in natural language understanding. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31:2947–2962.
- Yuhan Liu, Yuxuan Liu, Xiaoqing Zhang, Xiuying Chen, and Rui Yan. 2025. The truth becomes clearer through debate! multi-agent systems with large language models unmask fake news. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 504–514.
- Daze Lu, Vanessa Jiang, Shue Wang, Ming Zhu, and Shengjie Zhai. 2026. Enhancing medical question answering via a multi-agent debate framework. In *2026 IEEE 16th Annual Computing and Communication Workshop and Conference (CCWC)*, pages 118–124. IEEE.
- Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36:53728–53741.
- Bianca Raimondi, Francesco Pivi, Davide Evangelista, and Maurizio Gabbriellini. 2026. The compmath-mcq dataset: Are llms ready for higher-level math? *arXiv preprint arXiv:2603.03334*.

- Ankur Samanta, Akshayaa Magesh, Runzhe Wu, Ayush Jain, Youliang Yu, Daniel Jiang, Boris Vidolov, Paul Sajda, Yonathan Efroni, and Kaveh Hassani. 2025. Self-improvement of language models by post-training on multi-agent debate. *arXiv preprint arXiv:2509.15172*.
- Sina Shool, Sara Adimi, Reza Saboori Amleshi, Ehsan Bitaraf, Reza Golpira, and Mahmood Tara. 2025. A systematic review of large language model (llm) evaluations in clinical medicine. *BMC Medical Informatics and Decision Making*, 25(1):117.
- Andries Petrus Smit, Paul Duckworth, Nathan Grinsztajn, Kale-ab Tessera, Thomas D Barrett, and Arnú Pretorius. 2023. Are we going mad? benchmarking multi-agent debate between language models for medical q&a. In *Deep Generative Models for Health Workshop NeurIPS 2023*.
- Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Shoeb, Abubakar Abid, Adam Fisch, Adam R Brown, Adam Santoro, Aditya Gupta, Adri Garriga-Alonso, and 1 others. 2023. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Transactions on machine learning research*.
- Zhengyang Tang, Xingxing Zhang, Benyou Wang, and Furu Wei. 2024. Mathscale: scaling instruction tuning for mathematical reasoning. In *Proceedings of the 41st International Conference on Machine Learning*, pages 47885–47900.
- Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V Le, Ed H Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations*.
- Haolun Wu, Zhenkun Li, and Lingyao Li. 2025. Can llm agents really debate? a controlled study of multi-agent debate in logical reasoning. *arXiv preprint arXiv:2511.07784*.
- Andrea Wynn, Harsh Satija, and Gillian Hadfield. 2025. Talk isn’t always cheap: Understanding failure modes in multi-agent debate. *arXiv preprint arXiv:2509.05396*.
- Kai Xiong, Xiao Ding, Yixin Cao, Ting Liu, and Bing Qin. 2023. Examining inter-consistency of large language models collaboration: An in-depth analysis via debate. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 7572–7590.
- Kai Xiong, Xiao Ding, Ting Liu, Bing Qin, Dongliang Xu, Qing Yang, Hongtao Liu, and Yixin Cao. 2024. Meaningful learning: Enhancing abstract reasoning in large language models via generic fact guidance. *Advances in Neural Information Processing Systems*, 37:120501–120525.
- An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.
- Shaowei Zhang and Deyi Xiong. 2025. Debate4math: Multi-agent debate for fine-grained reasoning in math. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 16810–16824.
- Yiqun Zhang, Xiaocui Yang, Shi Feng, Daling Wang, Yifei Zhang, and Kaisong Song. 2026. Can llms beat humans in debating? a dynamic multi-agent framework for competitive debate. In *ICASSP 2026-2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 19377–19381. IEEE.
- Huaqin Zhao, Zhengliang Liu, Zihao Wu, Yiwei Li, Tianze Yang, Peng Shu, Shaochen Xu, Haixing Dai, Lin Zhao, Hanqi Jiang, and 1 others. 2024. Revolutionizing finance with llms: An overview of applications and insights. *arXiv preprint arXiv:2401.11641*.
- Yutian Zhao, Huimin Wang, Yefeng Zheng, and Xian Wu. 2025. A layered debating multi-agent system for similar disease diagnosis. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 539–549.
- Wanjun Zhong, Siyuan Wang, Duyu Tang, Zenan Xu, Daya Guo, Jiahai Wang, Jian Yin, Ming Zhou, and Nan Duan. 2021. Ar-lsat: Investigating analytical reasoning of text. *arXiv preprint arXiv:2104.06598*.
- Yan Zhou and Yanguang Chen. 2025. Adaptive heterogeneous multi-agent debate for enhanced educational and factual reasoning in large language models. *Journal of King Saud University Computer and Information Sciences*, 37(10):330.

A Prompt Templates

This appendix lists the prompt templates used in our method and baseline methods. For our method, we provide the initial debate prompt and the structured debate-turn prompt. For baselines, we implement task-specific prompt templates following the corresponding original methods, including Majority Voting, Self-Consistency, FORD, and MAD. Since these methods were originally proposed under different settings, we adapt their prompts to the same multiple-choice reasoning format. For fair comparison and stable answer extraction, all methods are required to output responses in a unified Analysis + Answer format.

In the following templates, {question} denotes the input question, {options} denotes the answer options, {allowed_labels} denotes the valid option labels, and {debate_history} denotes the previous debate context.

A.1 Initial Debate Prompt of Our Method

Our method treats the initial turn as a special case. The first agent independently analyzes the question and outputs its initial commitments and answer. The second agent also forms its own answer independently, and then compares it with the first agent’s response to optionally provide rebuttals.

A.1.1 First Agent Initial Prompt

System Prompt

You are the PRO debating agent. Your job is to answer a multiple-choice reasoning question. You should read the question and options carefully, reason step by step in natural language, and then explicitly state your COMMITMENTS and your final answer. When reasoning, you may implicitly consider Toulmin-style elements such as data, reasoning links, background rules, confidence, and possible counterconditions. You do not need to strictly follow a Toulmin format.

Commitments. You must explicitly write your commitments using the following four types:

1. Factual commitments: concrete facts from the question you accept as true.
2. Inferential commitments: inferences you accept.
3. Normative/rule commitments: general rules or principles you rely on.
4. Claim commitments: your stance on which option is correct.

Characteristics of commitments.

- They are publicly accepted assumptions that remain in force unless you retract them.
- They should be internally consistent.
- The opponent can later attack these commitments, so make them precise and well grounded.

- Multiple commitments must be separated by periods. At the very end, give your current chosen option in the following format:

Answer: (X)

where X must be exactly one of {allowed_labels}. If there is no clearly correct option, choose the one that you think is the most reasonable.

User Prompt

Question:

{question}

Options:

{options}

Please respond in the following format:

Analysis: <your free-form reasoning in natural language>

Commitments:

- Factual commitments: c1. c2. c3.
- Inferential commitments: i1. i2. i3.
- Normative/rule commitments: n1. n2. n3.
- Claim commitments: cl1. cl2.

You must separate multiple commitments with periods.

Answer: (X)

where X must be exactly one of {allowed_labels}. If there is no clearly correct option, choose the one that you think is the most reasonable.

A.1.2 Second Agent Initial Prompt

System Prompt

You are a debating agent participating in solving a multiple-choice reasoning problem.

Very important.

- When answering the question, you must reason independently.
- Do not treat the other agent’s analysis or answer as evidence.
- First, decide your own answer based only on the question and options.
- Only after you have formed your own view, you may compare it with the other agent and optionally rebut.

Your goals in each response are:

1. Think independently and decide which option you find most reasonable.
2. Write your own analysis based on the question.
3. Compare your answer with the other agent’s answer. If you agree, you may skip rebuttal. If you disagree, provide one or more rebuttals targeting the other agent’s arguments or commitments.
4. Output your updated commitments and your final answer.

Rebuttal guideline. If you choose to rebut, you may use one or more of the following attack types:

- factual attack;
- warrant attack;

- backing attack;
- counterexample;
- qualifier/condition attack.

Commitments. You must always output four types of commitments:

1. Factual commitments: concrete facts from the question you accept as true.
2. Inferential commitments: inferences you accept.
3. Normative/rule commitments: general rules or principles you rely on.
4. Claim commitments: your stance on which option is correct.

Multiple commitments must be separated by periods. At the end of your response, always provide:

Updated Commitments

Answer: (X)

where X must be exactly one of {allowed_labels}. If there is no clearly correct option, choose the one that you think is the most reasonable.

Effect: how this affects the plausibility of the other agent's claim

5. Updated Commitments.

- Factual commitments: c1. c2. c3.
- Inferential commitments: i1. i2. i3.
- Normative/rule commitments: n1. n2. n3.
- Claim commitments: c11. c12.

You must separate multiple commitments with periods.

6. Final Answer.

Answer: (X)

where X must be exactly one of {allowed_labels}. If there is no clearly correct option, choose the one that you think is the most reasonable.

User Prompt

Question:

{question}

Options:

{options}

Your response must contain the following parts:

1. **Analysis.** First, reason about the question and options independently. Do not copy or rely on any other agent's analysis as evidence. Write your own free-form reasoning based only on the question and options.
2. **Provisional Answer.** Based on your independent analysis, give a provisional answer in the format:

Answer: (X)

where X must be exactly one of {allowed_labels}.

3. **Other agent's answer for comparison and possible rebuttal.**

{first_agent_response}

4. **Rebuttal.** Compare your provisional answer and reasoning with the other agent's answer. If your answer is the same and you agree with the other agent's reasoning, write:

Rebuttal: I agree. No rebuttal.

Otherwise, provide one or more rebuttals in the following lightweight format:

Rebuttal 1:

Target: which part of the other agent's argument or which commitment you attack

Grounds: why this target is problematic

Attack Type: factual / warrant / backing / counterexample / qualifier/condition

A.2 Structured Debate-turn Prompt of Our Method

After the initial turn, each agent receives the question and the full debate history, and then produces a structured debate move. The prompt explicitly requires the model to defend its own challenged commitments, attack the opponent's commitments, update its commitment set, and provide a final answer.

System Prompt

You are now participating in a debate about a reasoning question. The debate has been transformed into an attack-defense process over commitment sets.

Commitments come in four types:

1. Factual commitments: concrete facts accepted from the question.
2. Inferential commitments: accepted inferences.
3. Normative/rule commitments: general rules or principles.
4. Claim commitments: stance on which answer option is correct.

You can respond with both defense and attack moves.

Defense.

- Correct misunderstanding of your commitments.
- Provide justification for disputed commitments.
- Retract or revise commitments if the opponent's attack is valid.

Attack.

- Factual attack: challenge the accuracy or completeness of the opponent's facts.
- Warrant attack: challenge the logical link between data and claim.
- Backing attack: question the general rule or background assumption.
- Counterexample: provide an example contradicting the opponent's claim.

- **Qualifier/condition attack:** show that required conditions do not hold.

Your answer must follow this structure:

1. **Defense.** If the opponent has attacked your commitments, list your defenses as Defense 1, Defense 2, etc. Each defense must contain:

- **Target:** which commitment or part of your argument you are defending.
- **Response:** how you correct, justify, or adjust this commitment.
- **Decision:** explicitly say keep, modify, or retract this commitment.

If no defense is needed, write: Defense: No defense needed.

2. **Attack.** List your attacks as Attack 1, Attack 2, etc. Each attack must contain:

- **Target:** which opponent commitment or argument you attack.
- **Grounds:** why this target is problematic.
- **Attack Type:** one of factual attack, warrant attack, backing attack, counterexample, qualifier/condition attack.
- **Effect on Opponent Claim:** how this attack changes the plausibility of the opponent's claim.

If you decide not to attack, write: Attack: No attack.

3. **Updated Commitments.** After defense and attack, list your updated commitments by type. Each type must be written in one line, with multiple commitments separated by periods:

- Factual commitments: c1. c2. c3.
- Inferential commitments: i1. i2. i3.
- Normative/rule commitments: n1. n2. n3.
- Claim commitments: c11. c12.

4. **Final Answer.** At the very end, give your current chosen option in the format:

Answer: (X)

where X must be exactly one of {allowed_labels}. If there is no clearly correct option, choose the one that you think is the most reasonable.

User Prompt

Question:
{question}

Debate History:
{debate_history}

Now produce your next move.
Follow this exact reply format:

1. **Defense.** Defense k, if any:
 - **Target:** which commitment or part of your argument you are defending.
 - **Response:** how you correct, justify, or adjust this commitment.
 - **Decision:** explicitly say keep, modify, or retract this commitment.
2. **Attack.** Attack k, if any:

- **Target:** which opponent commitment or argument you attack.
- **Grounds:** why this target is problematic.
- **Attack Type:** one of factual attack, warrant attack, backing attack, counterexample, qualifier/condition attack.
- **Effect on Opponent Claim:** how this attack changes the plausibility of the opponent's claim.

If you decide not to attack, explicitly write: Attack: No attack.

3. **Updated Commitments.**

- Factual commitments: c1. c2. c3.
- Inferential commitments: i1. i2. i3.
- Normative/rule commitments: n1. n2. n3.
- Claim commitments: c11. c12.

4. **Final Answer.**

Answer: (X)

where X must be exactly one of {allowed_labels}. If there is no clearly correct option, choose the one that you think is the most reasonable.

A.3 Majority Voting Prompt

The Majority Voting baseline samples multiple independent answers using the following prompt and selects the final answer by majority vote. The prompt follows the standard independent-answer setting and uses the same output format as other methods.

System Prompt

You are an AI agent answering a multiple-choice question independently. Select the best answer based only on the given question and options.

User Prompt

Answer the following multiple-choice question.

Question:
{question}

Options:
{options}

Give a brief analysis if needed, then state your final answer at the very end.

Use exactly the following format:

Analysis: (brief reason for your choice)

Answer: (Your selected option letter, e.g., (A))

A.4 Self-Consistency Prompt

The Self-Consistency baseline follows the idea of sampling multiple reasoning paths and aggregating the final answers by majority vote. We adapt it to our multiple-choice reasoning setting and use the same answer extraction format.

System Prompt

You are a reasoning model. For each question, generate a clear reasoning process before giving the final answer. The final answer must be one option letter.

User Prompt

Answer the following multiple-choice question.

Question:
{question}

Options:
{options}

Let's think step by step. Analyze the problem carefully, compare the options, and derive the answer through reasoning.

Use exactly the following format:

Analysis: (your step-by-step reasoning)

Answer: (Your selected option letter, e.g., (A))

A.5 MAD-style Debate Prompt

We construct the MAD-style debate prompt by adapting the debate setting of *Encouraging Divergent Thinking in Large Language Models through Multi-Agent Debate* to our multiple-choice reasoning tasks. The prompt is task-specific and uses the same unified answer format for stable answer extraction.

A.5.1 Initial MAD-style Prompt

System Prompt

You are in a debate now. The other agent's opinion is not always true, and you can ignore any incorrect part of it. You may refer to the other agent's opinion to revise your choice or defend your own. Use your general knowledge and understanding to read the other agent's opinion carefully and compare it with your own opinion based on the question. There should be a more plausible answer among the choices.

User Prompt

Question:
{question}

Options:
{options}

Please defend your opinion or give in to the other agent's opinion. Give your final explanation. Do not copy the other agent's opinion.

Use the following format:

Analysis: (your concise reasoning)

Answer: (X)

where X must be exactly one of the valid option labels. If there is no clearly correct option, choose the one that you think is the most reasonable.

A.5.2 Later MAD-style Debate Prompt

For later debate turns, we append the previous debate history before the current instruction. The model is then asked to continue the debate and output a concise response in the unified format.

System Prompt

You are in a debate now. The other agent's opinion is not always true, and you can ignore any incorrect part of it. You may refer to the other agent's opinion to revise your choice or defend your own. Use your general knowledge and understanding to read the other agent's opinion carefully and compare it with your own opinion based on the question. There should be a more plausible answer among the choices.

User Prompt

Question:
{question}

Options:
{options}

Debate History:
{debate_history}

Please continue the debate based on the previous context. You may defend your own opinion or revise your answer if the other agent's reasoning is stronger.

Use the following format:

Analysis: (concise reasoning)

Answer: (Your selected option letter, e.g., (A))

A.6 FORD-style Debate Prompt

We construct the FORD-style baseline by adapting the formal debate setting of FORD to our multiple-choice reasoning tasks. The initial turn asks each model to independently answer the question, and later turns ask the model to evaluate the opponent's reasoning and defend its own answer.

A.6.1 Initial FORD-style Prompt

System Prompt

You are a deductive reasoning model. Your task is to analyze the question carefully, consider all options, and choose the most logically correct answer.

User Prompt

You need to answer the following question and choose one of the following options as your answer.

Question:
{question}

Options:
{options}

You must first write a short reasoning section labeled Analysis, explaining your reasoning process.

Use the following format:

Analysis: (your reasoning)

Answer: (X)

where X must be exactly one of the valid option labels. If there is no clearly correct option, choose the one that you think is the most reasonable.

A.6.2 Later FORD-style Debate Prompt

For later debate turns, we append the previous debate history to the input. The model is asked to analyze the opponent’s argument, defend its own answer when appropriate, and provide a final answer in the unified format.

System Prompt

You are a reasoning model participating in a debate with another model.

Both agents propose answers based on the question and options. The goal is to defend your choice robustly and evaluate the reasoning behind the opponent’s choice.

During the debate:

- Analyze the opponent’s argument before crafting your own response.
- Provide a clear explanation for your answer.
- Focus on logical reasoning and common sense.
- Do not settle for weak or poorly reasoned arguments.
- Revise your answer only when the opponent’s reasoning is stronger.

User Prompt

Question:

{question}

Options:

{options}

Debate History:

{debate_history}

Please continue the debate based on the previous context. Evaluate the opponent’s argument, defend or revise your own answer, and give your final response.

Use the following format:

Analysis: (concise reasoning)

Answer: (Your selected option letter, e.g., (A))

B Data Construction and Filtering Details

This appendix provides additional details on how we construct, filter, and transform debate trajectories for supervised imitation learning. The filtering process is designed to retain trajectories that are not only outcome-correct, but also structurally valid and argumentatively meaningful.

B.1 Filtering Rules

Given a generated debate trajectory, we apply a set of rule-based filters to select high-quality training data. The filtering process considers five aspects: final-answer correctness, structured format validity, attack-defense grounding, final commitment consistency, and contribution quality.

First, we require the final answer of the trajectory to be correct, so that the retained trajectories provide useful supervision for task solving. Second, we check whether the generated responses follow the required structured format. Specifically, each response should contain commitments or updated commitments, and each non-initial debate response should contain both defense and attack blocks. A defense block is considered valid if it contains the fields Defense, Target, Response, and Decision. An attack block is considered valid if it contains the fields Attack, Target, Grounds, Attack Type, and Effect on Opponent Claim.

Third, we check whether attacks and defenses are grounded in relevant argumentative content. We first extract the target text from each defense or attack unit. Defense targets are compared with the model’s earlier commitments, while attack targets are compared with the opponent’s previous commitments. We compute semantic similarity between each target and the corresponding commitment set, and use the best-matched commitment to score grounding quality. The average best-match semantic similarity should be at least 0.65.

Fourth, we check final commitment consistency. This constraint is applied only when the last two responses produce the same answer. In this case, their commitment sets should have a semantic similarity of at least 0.8. The similarity between two commitment sets is computed by bidirectional best-match averaging. If the last two answers are different, this constraint is not triggered.

Finally, we compute a contribution score to retain trajectories with useful belief-transition behavior. The score is computed from non-initial responses. If the current answer is wrong, the contribution score is -1 . If the current answer is correct and the previous two answers are both correct, the score is 0.1, which represents stable maintenance of a correct belief. If the current answer is correct, the previous answer is correct, the earlier answer is wrong, and the current turn contains a revision signal such as retract, modify, revised, or withdraw, the score is 1. If the current answer

is correct, the previous answer is wrong, and the current turn contains an attack with an explicit target, the score is also 1. All other cases receive a score of 0.

The final contribution score of a trajectory is the average score over its non-initial responses. We retain trajectories whose average contribution score is larger than 0.

B.2 Statistics of Constructed Debate Data

Before filtering, we first generate structured debate trajectories on the training split. Table B.1 reports the accuracy of different agents or debate methods during this stage. *Size* denotes the number of training instances in each dataset.

We then apply the filtering rules described in Section B.1. Table B.2 summarizes the number of retained trajectories under different filtering stages. *Both Correct* $\rightarrow$ *Correct* denotes trajectories where both agents are initially correct and the final answer remains correct. *Both Wrong* $\rightarrow$ *Correct* denotes trajectories where both agents are initially wrong but the final answer becomes correct. *Format-valid* denotes trajectories that pass the structural and consistency filters, including format validity, attack-defense grounding, and final commitment consistency. *High-quality* denotes the final retained trajectories used for supervised imitation learning. Specifically, after obtaining the format-valid trajectories, we compute a contribution score for each non-initial response from context[2] onward, average these scores within each trajectory, and retain only trajectories whose average contribution score is larger than 0.

B.3 Turn-level SFT Sample Construction

After filtering, we convert each high-quality debate trajectory into turn-level supervised fine-tuning examples. Each trajectory contains the original question, answer options, and a sequence of debate responses. We expand each trajectory by treating each target-agent turn as one training example.

For Qwen, even-indexed turns are used as target responses. For Llama, odd-indexed turns are used as target responses. Thus, each model is trained only on its own utterances, while the other agent’s utterances are treated as dialogue context.

For each training example, the input consists of the question, answer options, and all previous debate turns before the current response. The target output is the current response of the target agent.

Formally, for a trajectory with context sequence

$$H = (u^0, u^1, \dots, u^T), \quad (19)$$

we construct a training pair for a target-agent turn t as

$$s^t = (x, O, u^0, \dots, u^{t-1}), \quad y^t = u^t, \quad (20)$$

where x is the question, O denotes the answer options, s^t is the input context, and y^t is the supervised target response.

During training, all tokens in the input context are masked with -100 , and the loss is computed only on the target response tokens. This ensures that the model learns to generate its own structured debate move based on the previous debate history, rather than being optimized to reproduce the entire dialogue context.

C Reward and Metric Details

This appendix provides implementation details for the reward design used in Debate Internalization and the rule-based RES computation. The main text presents the overall reward formulation, while this appendix specifies the concrete scoring rules used in our implementation.

C.1 Utterance-level Reward Rules

For each generated utterance, we compute an utterance-level reward from five components:

$$r_{\text{utt}} = w_1 r_{\text{ans}} + w_2 r_{\text{reason}} + w_3 r_{\text{fmt}} + w_4 r_{\text{cons}} + w_5 r_{\text{info}}. \quad (21)$$

The weights are set to $w_1 = 0.7$, $w_2 = 0.5$, $w_3 = 0.12$, $w_4 = 0.15$, and $w_5 = 0.1$. The five components are summarized in Table C.1.

The reasoning-quality reward is computed from the correctness pattern of the model’s previous answer, the opponent’s previous answer, and the current answer. For a non-initial turn t , the model’s previous answer is taken from turn $t - 2$, and the opponent’s previous answer is taken from turn $t - 1$.

The scoring rules are category-aware. When both the model and the opponent were previously correct, the desired behavior is to keep the correct belief. In this case, a current correct answer receives $0.8 + 0.5 \times \text{conf}$, where conf is the model’s confidence in the correct answer, while a current wrong answer receives 0.05. When the model was previously correct but the opponent was wrong, the model should resist invalid compromise. Keeping

Dataset	Size	Llama	Qwen	MAD	Our Method
α NLI	1006	73.36	77.13	77.83	77.99
MMLU	1300	67.54	79.69	75.00	79.92
LogiQA2	1140	51.05	74.21	62.37	77.19
e-CARE	1622	77.44	81.38	77.99	81.20
BIG-bench	1200	43.25	77.92	77.50	82.17
AR-LSAT	1585	24.04	71.67	62.52	78.62

Table B.1: Accuracy of generated debate trajectories before filtering.

Dataset	Both Correct → Correct	Both Wrong → Correct	Format-valid	High-quality
α NLI	670	78	550	539
BIG-bench	954	13	757	633
MMLU	902	92	771	766
LogiQA2	773	53	670	663
AR-LSAT	793	39	371	344
e-CARE	1123	127	930	918
Total	5215	402	4049	3863

Table B.2: Statistics of filtered debate trajectories.

its own correct answer receives 1.55, switching to the opponent’s wrong answer receives 0.10, and other answers receive 0.22. When the model was previously wrong but the opponent was correct, the desired behavior is to accept the stronger argument. Switching to the opponent’s correct answer receives 1.78, keeping its own wrong answer receives 0.30, and other answers receive 0.42. When both were previously wrong, the model is encouraged to recover from wrong beliefs: a current correct answer receives 1.55, repeating a previous wrong answer receives 0.52, and choosing another wrong answer receives 0.64.

The format reward is normalized to $[0, 1]$. For initial responses, it checks whether the model provides an extractable answer and commitments. For later debate turns, it additionally checks whether the response contains attack or defense structures. A small penalty is applied when the response repeatedly outputs multiple Answer : fields.

The consistency reward checks whether the current defense targets the model’s own earlier commitments and whether the current attack targets the opponent’s previous commitments. We compute the best semantic similarity between each target and the corresponding commitment set. If the average similarity is at least 0.7, $r_{\text{cons}} = 1$; otherwise, $r_{\text{cons}} = 0$.

The new-information reward is computed with text-based heuristics. It rewards expressions that introduce new examples, conditions, qualifica-

tions, context, or alternative explanations, such as for example, not necessarily, depends on, without specifying, and could still. Repeated cue phrases are down-weighted to reduce template-like reward hacking.

C.2 Trajectory-level Reward Rules

In addition to utterance-level rewards, we compute a dialogue-level reward to evaluate whether the whole debate produces a desirable belief transition. The reward depends on the initial correctness state, final correctness, and whether the two agents reach agreement.

The dialogue-level reward is defined over four initial states. In Both Correct (BC), both agents are initially correct. If the final answer remains correct, the reward is 1.35 when the agents reach consensus and 1.20 when they do not; if the final answer is wrong, the reward is 0. In Trainable Correct (TC), the trainable agent is initially correct while the opponent is wrong. A correct final answer receives 1.65 with consensus and 1.45 without consensus, while a wrong final answer receives 0.05 with consensus and 0.10 without consensus. In Opponent Correct (OC), the trainable agent is initially wrong while the opponent is correct. A correct final answer receives 1.85 with consensus and 1.65 without consensus, while a wrong final answer receives 0.20 with consensus and 0.30 without consensus. In Both Wrong (BW), both agents are initially wrong. A correct final answer receives

Component	Meaning	Rule
r_{ans}	Answer correctness	1 if the extracted answer matches the ground-truth answer, otherwise 0.
r_{reason}	Debate reasoning quality	Rewards proper belief maintenance, rebuttal, compromise, or recovery according to the previous two answers.
r_{fmt}	Format validity	Checks whether the response contains the required answer, commitments, and attack/defense structure.
r_{cons}	Argumentative consistency	Checks whether attack/defense targets are semantically grounded in the corresponding commitment sets.
r_{info}	New information	Rewards useful new conditions, counterexamples, qualifications, or alternative reasoning paths.

Table C.1: Utterance-level reward components.

2.00 with consensus and 1.80 without consensus, while a wrong final answer receives 0.45 with consensus and 0.55 without consensus.

For each rollout τ_g , we also compute the mean utterance-level reward over the trainable agent’s turns:

$$\bar{r}_{\text{train}}(\tau_g) = \frac{1}{|T_p|} \sum_{t \in T_p} r_{\text{utt}}(u_{p,g}^t), \quad (22)$$

where T_p denotes the set of trainable-agent turns. The final rollout score is:

$$s(\tau_g) = 0.7 \cdot r_{\text{traj}}(\tau_g) + 0.3 \cdot \bar{r}_{\text{train}}(\tau_g). \quad (23)$$

For group-relative advantage estimation, mixed groups use standard normalization. When all rollouts in a group are correct, we add a pseudo-negative reference score $s^- = 1.22$ to compute the group statistics. When all rollouts are wrong, we add a pseudo-positive reference score $s^+ = 0.60$. These pseudo references are used only for computing the mean and standard deviation; they are not used as real training samples.

The rollout-level advantage determines whether a trajectory should be encouraged or discouraged. For each trainable-agent utterance, we further multiply the rollout advantage by its local utterance score, so that higher-quality utterances receive stronger credit during policy optimization.

C.3 Rule-based RES Scoring Details

Since BGS is fully defined in the main text, this section focuses on the implementation of Rebuttal Engagement Score (RES). RES is computed only on trainable-agent response turns, excluding the initial answer. In our two-agent setting, this corresponds to Qwen-side later debate turns.

For each selected turn, we extract attack and defense units. Each unit contains its type, target, response, and decision when available. If a response

does not contain explicit attack or defense fields but includes a rebuttal section, we treat it as a weakly structured attack unit.

We compute three core scores: target hit (TH), uptake response (UR), and focus continuity (FC). The similarity function uses rule-based matching by default, combining token-level overlap and character n-gram overlap. A continuous similarity score is mapped to a discrete score as follows: a similarity score of at least 0.42 is mapped to 1.0; a score between 0.20 and 0.42 is mapped to 0.5; and a score below 0.20 is mapped to 0.0.

TH measures whether the current attack or defense targets the opponent’s previous commitments or core arguments. If an explicit target is available, we compare it with the previous commitments and assign a score using the similarity mapping above. If no explicit target is available but the response explicitly refers to the opponent or the other agent, we assign a weak target-hit score of 0.5.

UR measures whether the model gives a substantive response after hitting the target. It combines target hit, response substance, and a small bonus for explicit decisions:

$$\text{UR} = \text{clip}_{[0,1]} (0.45 \cdot \text{TH} + 0.55 \cdot \text{Substance} + \text{DecisionBonus}). \quad (24)$$

The substance score depends on response length and argumentative cues such as because, therefore, however, weaken, support, contradict, revise, claim, and reasoning.

FC measures whether the current turn continues the same disputed point rather than shifting to an unrelated topic. We compare the current target or response with previous targets, responses, and commitments, and map the maximum similarity using the same threshold rule.

We also compute an argument-unit quantity

score:

$$\text{ADQ} = \min\left(\frac{N_{\text{unit}}}{E_{\text{unit}}}, 1\right), \quad (25)$$

where N_{unit} is the number of extracted attack or defense units, and $E_{\text{unit}} = 2$ is the expected number of argument units per response turn. The final turn-level RES is:

$$\text{RES} = (1-\eta) \cdot \frac{\text{TH} + \text{UR} + \text{FC}}{3} + \eta \cdot \text{ADQ}, \quad (26)$$

where $\eta = 0.15$. The final dataset-level RES is obtained by averaging over all valid trainable-agent response turns. The whole RES computation is rule-based and does not rely on an LLM judge.

D Responsible Use and Reproducibility Details

D.1 Artifact Licenses, Intended Use, and Data Risks

We use six publicly available multiple-choice reasoning benchmarks: α NLI, MMLU, LogiQA2, e-CARE, BIG-bench, and AR-LSAT. We follow the licenses or terms reported by the original public repositories or dataset cards. The datasets are used only for research-oriented training and evaluation, and our use is consistent with their intended purpose as reasoning and language understanding benchmarks.

Our work creates derived artifacts, including generated debate trajectories, filtered training data, evaluation outputs, and implementation code. We plan to release the code and derived data artifacts when permitted by the original dataset licenses and terms. Full model checkpoints are not released at this stage, but may be released separately subject to the original model licenses and hosting constraints.

We do not collect new user data, or introduce additional personally identifying information. Since the data are drawn from public benchmarks, the original datasets may still contain sensitive, biased, or potentially offensive content inherited from their sources. We do not add such content during data construction, and all generated trajectories are used only for research evaluation and model training. Before use, we reviewed the documentation and dataset cards of all benchmarks for known privacy and content risks. We also checked the constructed training artifacts to ensure that our data-generation pipeline did not introduce additional personally identifying information; no newly collected personal information is included, and therefore no additional anonymization was required.

Dataset	Train Size	Test Size
α NLI	1006	250
MMLU	1300	250
LogiQA2	1140	250
e-CARE	1622	250
BIG-bench	1200	250
AR-LSAT	1585	250
Total	7853	1500

Table D.1: Dataset statistics used in training and evaluation.

Stage	GPUs	Hours	GPU Hours
Qwen SFT	4	10	40
Llama SFT	4	10	40
RL training	4	75	300
Evaluation	1	70	70
Total	13	165	450

Table D.2: Approximate computational budget.

D.2 Benchmark and Dataset Statistics

Table D.1 summarizes the benchmark-level data statistics used in our experiments. The training sizes correspond to randomly sampled subsets from the original training sets. For evaluation, we use 250 test instances from each benchmark, resulting in 1,500 test instances in total. We do not use a separate development set.

D.3 Computational Budget

All main experiments are conducted on NVIDIA A100-80GB GPUs. For supervised imitation learning, we fine-tune Qwen3-4B-Instruct-2507 and Meta-Llama-3-8B-Instruct separately. Each SFT run uses 4 A100-80GB GPUs for approximately 10 hours. Debate Internalization is trained on 4 A100-80GB GPUs for approximately 75 hours and is evaluated at checkpoint-1200. A full evaluation run of our method requires approximately 70 A100-80GB GPU hours. The total computational budget is therefore approximately 450 A100-80GB GPU hours.

D.4 Implementation Packages and Parameters

Our implementation is based on PyTorch and HuggingFace Transformers. The SFT stage uses the HuggingFace Trainer interface with AutoTokenizer, AutoModelForCausalLM, and TrainingArguments. The RL stage uses a custom implementation built on PyTorch and Transformers. For semantic matching in filtering and reward com-

putation, we use sentence-transformers with bge-base-en-v1.5 when available; otherwise, the implementation falls back to token-overlap matching. Evaluation metrics are computed with rule-based extraction scripts and do not rely on an LLM judge.

The main hyperparameters are summarized below. For SFT, we use a learning rate of 1×10^{-5} , batch size 4, 3 epochs, warmup steps 50, and maximum sequence length 5120. For Debate Internalization, we use group size 4, maximum turns 6, temperature 0.8, top- p 0.95, maximum new tokens per turn 800, learning rate 5×10^{-6} , KL coefficient 0.02, clip threshold 0.2, update epochs 2, mini-batch utterances 4, gradient accumulation 2, and checkpoint-1200 for final evaluation. During debate evaluation, we use maximum turns 6, temperature 0, top- p 0.9, maximum new tokens 4096, and stop-on-consensus decoding.

For rule-based filtering and scoring, the attack-defense grounding threshold is 0.65, the final commitment consistency threshold is 0.8, and the consistency reward threshold is 0.7. For RES, the high and middle similarity thresholds are 0.42 and 0.20, respectively; the expected number of argument units is 2, and the argument-unit weight is $\eta = 0.15$.

D.5 Human Evaluation Details

Evaluation Setup. For the human evaluation in Section 5.4, we randomly sample 20 questions from each of the six evaluation benchmarks, resulting in 120 questions in total. For each question, we collect the debate trajectories produced by Debate Internalization (DI) and GRPO. The method identities are hidden from the annotators during evaluation. Three annotators independently evaluate each comparison.

Annotators are provided with the original question, answer options, and two anonymized debate trajectories, denoted as Debate A and Debate B. The ground-truth answer is not provided to the annotators. The ordering of the two debate trajectories is anonymized so that annotators do not know which method produces either trajectory.

Annotation Instructions. For each pair of debate trajectories, annotators evaluate the following three aspects:

1. **Overall debate quality:** Which debate exhibits better overall debate quality?
2. **Rebuttal engagement:** Which debate more

directly and substantively engages with the opponent’s arguments?

3. **Belief update quality:** Which debate maintains or updates its stance more reasonably according to the arguments presented during the interaction?

For each aspect, annotators perform two types of assessment. First, they make a pairwise preference judgment and explicitly select either Debate A or Debate B as better; ties are not allowed. Second, they independently assign a quality score from 1 to 5 to each debate, where a higher score indicates better performance on the corresponding aspect. The two debates in the same comparison are not allowed to receive the same score.

The first aspect is used to measure direct human preference between DI and GRPO. The second aspect is used to validate the Rebuttal Engagement Score (RES), while the third aspect is used to validate the Belief Governance Score (BGS). Since annotators are not given the ground-truth answer, their judgments of belief update quality are based on whether stance maintenance or revision is reasonable given the arguments presented during the debate.

Agreement Computation. For each comparison and each evaluation aspect, the sample-level human preference is determined by majority vote among the three annotators. Since each annotator must choose either Debate A or Debate B, every comparison yields a unique majority preference.

For the direct DI-versus-GRPO comparison, we use the majority judgment on overall debate quality and report the proportion of samples for which DI is preferred.

For RES and BGS validation, we compare the pairwise ranking induced by the corresponding automatic metric with the majority human preference on rebuttal engagement and belief update quality, respectively. A comparison is counted as agreement when the automatic metric assigns a higher score to the same debate preferred by the majority of human annotators.

We consider two settings. In the same-question setting, DI and GRPO debate trajectories generated for the same question are compared. In the cross-question setting, we additionally construct 60 pairs of different questions within DI. The same three annotators evaluate these pairs using the same rebuttal-engagement and belief-update criteria, and

we measure whether the human and automatic metric rankings agree.

Annotator Recruitment and Compensation.

We recruit three annotators with backgrounds in LLM reasoning and debate research. Each annotator is paid \$10 per hour, while the minimum hourly wage in our country is \$5. Thus, the compensation provided for the annotation task is above the local minimum wage.

Consent and Privacy. Before annotation, all annotators are informed that their judgments will be used for research and publication purposes, and all annotators agree to the use of their annotations. The reported results contain only aggregated annotation judgments and do not include personally identifying or sensitive information about the annotators.

D.6 Use of AI Assistants

AI assistants were used only for language polishing and grammatical refinement during manuscript preparation. They were not used to generate experimental results, perform evaluations, make scientific conclusions, or replace the authors’ analysis. All AI-assisted revisions were reviewed and verified by the authors, who take full responsibility for the final content.