When Debate Fails: An Empirical Study of Incentive Misalignment in Prover–Estimator Games

Hannah Guan, Charles Hou
Harvard University
John A. Paulson School of Engineering and Applied Sciences
{hguan, charleshou}@college.harvard.edu

Abstract

Debate has been proposed as a mechanism for eliciting truthful reasoning from powerful learning agents by structuring interaction as a game between competing provers evaluated by an estimator. While theoretically appealing, the practical effectiveness of such incentive-based mechanisms remains underexplored. We present the first end-to-end empirical instantiation of this protocol using reinforcement learning agents with large language model on verifiable long-context reasoning tasks. Despite careful alignment of rewards with theoretical assumptions, we find that the debate mechanism consistently fails to elicit truthful or robust reasoning. Instead, agents converge to degenerate strategies that exploit estimator weaknesses or collapse into non-informative equilibria. Through controlled experiments and representation analysis using PCA and t-SNE, we identify a fundamental incentive–optimization mismatch: equilibrium incentives do not reliably translate into learnable equilibria under gradient-based optimization. Our results highlight a gap between game-theoretic guarantees and learned agent behavior, raising concerns about the reliability of debate-based alignment mechanisms in practice. We conclude with implications for incentive design and future directions for empirically grounded mechanism design in multi-agent systems.

Introduction

As large language models (LLMs) become increasingly capable of solving complex reasoning and decision-making tasks, ensuring that their outputs are truthful, reliable, and verifiable becomes a central challenge for AI alignment. In many high-stakes domains such as mathematics, science, law, and policy, LLMs may produce answers that appear plausible yet are difficult or prohibitively expensive for humans to verify directly. This asymmetry between generation and verification motivates the paradigm of *scalable oversight* (Engels et al. 2025), in which weaker human judges are assisted by AI systems. Scalable oversight seeks to simplify verification without sacrificing correctness guarantees.

One prominent proposal in this direction is *multi-agent debate*, where two competing AI agents argue for opposing answers to a question, recursively decomposing their claims into subclaims until a less capable judge can reliably determine which side is correct. The key insight is that while verifying a complete solution may be intractable, judging between two competing decompositions of the same problem can be significantly easier. Irving et al. (2018) (Irving, Christiano, and Amodei 2018) formalize this intuition and show that debate protocols can, in principle, allow polynomial-time verification of PSPACE-complete problems, suggesting that debate could enable oversight of reasoning far beyond direct human capabilities.

However, Barnes (2020) (Barnes 2020) highlights a core failure mode of traditional recursive debate: the *obfuscated arguments problem*. A dishonest agent may generate decompositions in which the flaw is buried in a subclaim that is easy to produce but extremely hard, or even intractable, to refute. This exploits the recursive nature of debate: if verification requires solving a harder problem than the original claim, the oversight mechanism breaks down. For example, a prover could decompose a simple arithmetic claim into subclaims involving primality testing or discrete logarithms that make verification intractable (Bi et al. 2025).

Brown-Cohen et al. (2025) (Brown-Cohen, Irving, and Piliouras 2025) propose the *prover–estimator debate* framework as a response to obfuscated arguments. Instead of having a challenger select one subclaim to attack, they introduce an *estimator* that assigns a probability to each subclaim being correct. The prover then chooses to recurse on the subclaim where it believes the estimator's probability is most wrong. Intuitively, because the estimator must expose its probabilistic beliefs over *all* subclaims, a dishonest prover cannot safely hide behind a single obfuscated branch: any systematic miscalibration becomes a source of exploitable reward for the other player.

Despite this theoretical progress, it remains unclear whether prover–estimator debate can be implemented effectively with real LLMs, finite compute, and noisy training objectives. In particular, the theory assumes that agents can play approximate best responses and that incentives are faithfully reflected in the learning signal. Buhl et al. (2025) (Buhl et al. 2025) explicitly suggest taking a pre-trained LLM and fine-tuning it via a debate process that satisfies the theoretical requirements, but do not instantiate or empirically evaluate such a system.

In this paper, we show that when prover–estimator de-

 Both authors contributed equally to this work.

bate is instantiated using finite-capacity language models trained via reinforcement learning (RL), these assumptions systematically break down. Our results suggest a fundamental mismatch between the game-theoretic incentives of prover–estimator debate and the optimization dynamics induced by RL fine-tuning. This mismatch implies that incentive-compatible equilibria in debate-based mechanisms do not automatically translate into learnable equilibria. This calls into question the reliability of such mechanisms when deployed with gradient-trained agents and motivates the need for designs that explicitly account for learning dynamics.

This project makes the following contributions:

- **A practical instantiation of prover-estimator debate.** We develop a complete reinforcement learning environment that operationalizes prover-estimator debate with LLMs. Our framework includes recursive debate trees, explicit prover and estimator roles, reward shaping aligned with the theoretical objectives, and termination at verifiable leaf nodes.
- **The first end-to-end empirical evaluation under gradient–based optimization.** We evaluate prover-estimator debate on challenging long-context multiple-choice reasoning benchmarks with known ground-truth answers but solutions require multi-step reasoning. This enables controlled measurement of estimator calibration, prover adaptation, decomposition quality, and overall debate behavior under RL fine-tuning.
- **Identification and analysis of an incentive–optimization mismatch.** We show that the failure we observe is not merely due to insufficient scale. By mapping empirical failures to violated theoretical assumptions, we diagnose a structural mismatch between game-theoretic guarantees and the optimization dynamics of gradient-trained language models.

Debate Protocols as Incentive Mechanisms

Scalable Oversight and AI Alignment

The fundamental challenge of *scalable oversight* – supervising AI systems on tasks that exceed human capability – has been identified as central to safe AI development (Amodei et al., 2016 (Amodei et al. 2016); Christiano et al., 2018 (Christiano, Shlegeris, and Amodei 2018)). Traditional supervised learning assumes human feedback on all training examples, but this breaks down when AI systems solve problems humans cannot easily verify.

Christiano et al., 2018 (Christiano, Shlegeris, and Amodei 2018) addresses this by recursively decomposing hard tasks into easier subtasks that humans can evaluate, building up capabilities through composition. However, amplification relies on humans to perform the decomposition, which may not scale to superhuman performance. Leike et al., 2018 (Leike et al. 2018) trains reward models recursively, using simpler reward models to supervise more complex ones. While promising, this approach still requires ground-truth labels at the base level and may propagate errors through the recursion. Irving et al., 2018 (Irving, Christiano, and Amodei 2018) offers an alternative: two AI agents compete to convince a judge of opposing answers, with the assumption that truth is easier to defend than falsehood. This shifts the burden from the human (who only judges arguments) to the AI agents (who generate and critique arguments). Our work implements a specific debate protocol designed to avoid the computational pitfalls that plague traditional debate.

Debate Protocols

Irving et al. (2018) (Irving, Christiano, and Amodei 2018) proved that debate between polynomial-time agents can enable a polynomial-time judge to verify solutions to any problem in PSPACE, assuming perfect play. This established debate as a theoretically powerful mechanism for scalable oversight by showing that verification complexity can be reduced through adversarial decomposition. However, the original protocol assumes that the judge can arbitrarily choose which subclaims to examine, an assumption that becomes problematic in realistic settings and enables obfuscation attacks when strategic agents exploit this freedom to hide errors in difficult-to-verify branches.

Several works have implemented debate systems with language models. Perez et al. (2019) (Perez et al. 2019) demonstrated debate on factual questions, showing that human judges preferred arguments from truthful debaters. Khan et al. (2021) (Khan et al. 2021) applied debate to question answering, finding that debate improved answer quality on multi-hop reasoning tasks. Michael et al. (2023) (Michael, Shlegeris, and Perez 2023) implemented debate for long-form question answering, demonstrating that debate can surface relevant information even when individual models are unreliable. However, these implementations focused on empirical performance rather than computational efficiency guarantees. None addressed the obfuscation problem or provided mechanisms to ensure honest provers are not disadvantaged by computationally hard decompositions.

Prover-Estimator Debate

Brown-Cohen et al. (2025) (Brown-Cohen, Irving, and Piliouras 2025) proposed the prover-estimator protocol specifically to avoid obfuscation while maintaining efficiency. The protocol involves three agents with distinct roles:

- Prover: An untrusted agent that makes claims and recursively decomposes them into simpler subclaims. The Prover aims to maximize reward by demonstrating when the Estimator's probability assignments are miscalibrated.
- Estimator: A trusted agent that assigns probabilities to claims without choosing which subclaims to examine. The Estimator aims to minimize reward loss through well-calibrated probability estimates.
- Oracle: A ground-truth verifier that evaluates atomic claims at the leaves of the recursion tree. The oracle is assumed to be correct but computationally limited (polynomial-time).

The key innovation is the asymmetry in control: the Prover generates subclaims and selects which branch to recurse on, while the Estimator must assign probabilities to all subclaims without controlling the recursion path.

The protocol satisfies two key properties:

- Completeness: If the Prover makes correct claims and generates valid, stable decompositions, the expected reward is positive: $\mathbb{E}[R] > 0$. This ensures that honest provers are not disadvantaged.
- Soundness: If the Prover makes incorrect claims or invalid decompositions, and the Estimator assigns well-calibrated probabilities, the expected reward is negative: $\mathbb{E}[R] < 0$. This ensures dishonest provers cannot benefit.

These guarantees hold when the following assumptions are true:

- Stability: The decomposition is s-supported, meaning small changes in subclaim probabilities don't drastically change the parent claim's truth value.
- A-provability: The claim can be proven via polynomial-depth decomposition into oracle-verifiable subclaims.
- Equilibrium: The Estimator plays a best-response strategy to the Prover's decomposition policy.

Intuitively, by letting both the prover and estimator influence how the debate recurses, prover-estimator debate makes it difficult for any single agent to win the debate by simply obfuscating their arguments. Our work provides the first complete empirical evaluation of this protocol.

Methodology

Task Definition

We study long-form multiple-choice reading comprehension as instantiated by the QuALITY benchmark. Each *episode* corresponds to a single reading comprehension instance consisting of:

- A passage p (often several thousand words).
- A question x about the passage.
- A finite answer set $\mathcal{A} = \{a_1, \ldots, a_K\}$ of multiple-choice options.
- A ground-truth answer $a^\star \in \mathcal{A}$, verified by human annotators.

For the purposes of debate, we frame each episode around an *answer claim*. Given a candidate answer $\tilde{a} \in \mathcal{A}$ (e.g., proposed by a base model or sampled from the options), we define a root claim:

$c^{(0)}$: "$\tilde{a}$ is the correct answer to question x about passage p."

A debate protocol then defines a sequence of messages between agents that recursively decompose and evaluate this claim, ultimately producing:

- A final binary verdict on $c^{(0)}$ (accept/reject the claim), and
- A corresponding chain of reasoning in the form of intermediate subclaims.

This formulation allows us to evaluate debate protocols in terms of standard multiple-choice accuracy, while the binary truth value of each intermediate claim provides supervision for training and assessing the prover and estimator.

Prover–Estimator Debate Framework

We instantiate prover-estimator debate as a sequential decision process over decomposition trees. At a high level, a debate proceeds as follows:

1. **Initialization.** The root claim $c^{(0)}$ asserts that a particular answer $a \in \mathcal{A}$ is correct for question x.
2. **Prover decomposition.** At step t, given the current claim $c^{(t)}$ and the debate history, the prover proposes a decomposition into subclaims:
$$c^{(t)} \to \{c_1^{(t)}, \ldots, c_{k_t}^{(t)}\}.$$
3. **Estimator probabilities.** The estimator observes the same context and outputs probabilities $p_i^{(t)} \in [0, 1]$ for each subclaim $c_i^{(t)}$ being correct, forming a categorical distribution or a vector of independent Bernoulli beliefs.
4. **Prover selection.** The prover then selects an index $j_t \in \{1, \ldots, k_t\}$ indicating which subclaim to recurse on. Intuitively, the prover aims to pick a subclaim where the estimator is most mistaken.
5. **Recursion.** The selected subclaim $c_{j_t}^{(t)}$ becomes the new current claim $c^{(t+1)}$, and the process repeats until termination.
6. **Termination.** Once a leaf claim is reached, i.e. the debate has decomposed $c^{(T)}$ into a form that is deemed simple, an oracle evaluates its correctness. This evaluation induces correctness labels for the path of subclaims and ultimately determines whether the top-level answer is correct.

In the idealized theory, the estimator is perfectly truthful and calibrated, and the prover is either honest or adversarial. Under stability conditions on the decomposition, Brown-Cohen et al. show that honest strategies are favored. In our empirical setting, both prover and estimator are parameterized LLM policies with finite capacity and are trained jointly via RL to optimize their respective objectives.

Model Details

Base models. For efficiency, we use small open-source instruction-tuned LLMs as base policies for both roles, starting with Llama-3.2-1B-Instruct. We initialize the prover and estimator from the same base checkpoint but maintain separate copies of the weights during fine-tuning, allowing the two roles to specialize.

Oracle verification. In the theoretical framework, debates terminate at leaves that are easy for humans to verify. Since we assume that any question in the QuALITY dataset would be easily answered by any human with basic reading comprehension skills, we implement the oracle as a state-of-the-art LLM – here, we use gpt-5-mini and Gemini Flash 2.0 via API. To reflect the stability requirement, we perform multiple independent "mini-checks" at the leaves and take a majority vote.

Experimental Setup

Baselines

We compare prover–estimator debate against two baselines, all trained from the same base model and evaluated under matched token and compute budgets. The simplest baseline is a single model that receives the question x and directly outputs an answer $\hat{a} \in \mathcal{A}$. This model can be fine-tuned using supervised learning on chain-of-thought data or via RL with a reward of $+1$ for correct answers and 0 otherwise. Direct-answer performance establishes a lower bound for what can be achieved without any multi-agent structure. The second baseline is a conventional two-agent debate protocol:

- A *prover* decomposes the top-level claim into subclaims at each step.
- A *challenger* selects one subclaim to attack, based on its own beliefs about where a flaw is most likely.
- The debate recurses on the selected subclaim until reaching a leaf.

At termination, an oracle evaluates the leaf claim, and the outcome is propagated back through the debate tree to determine which agent "won" the debate. The training objective encourages the prover to win when defending true claims and the challenger to win when attacking false ones, aligning incentives with correctness. This baseline isolates the contribution of the estimator's probabilistic predictions by matching the overall debate structure, including the decomposition tree, depth limits, and branching constraints, to those used in the prover–estimator protocol. As a result, any observed performance differences can be attributed primarily to the presence or absence of explicit probabilistic estimation rather than to structural differences in the debate setup.

Implementation Details

Training loop. All training is orchestrated by an alternating Stackelberg-style optimization between the prover (leader) and estimator (follower). For each training iteration (up to 100 iterations):

1. A batch of debate claims is sampled from QuALITY-derived questions.
2. For each instance in the batch, the trainer executes 3 full debate rollouts between the prover LLM (Alice) and estimator LLM (Bob). Each rollout proceeds until a maximum depth or termination condition is reached.
3. The estimator is updated *every* iteration as the follower. Using the most recent debate transcripts, it minimizes the Brier score so that its probability assignments better match oracle truth. We optimize the estimator for calibration because in theory, the estimator is supposed to be truthful to defend against dishonest provers.
4. The prover is updated periodically every 5 iterations. Using GRPO with a KL penalty toward the base model, the prover treats debate transcripts as RL trajectories and adjusts its policy to maximize the episode rewards.

Reward structure. The implemented reward signals follow the theoretical objectives but are computed directly from oracle-labeled transcripts:

- **Estimator reward.** For each subclaim c_i with oracle truth value $y_i \in \{0, 1\}$ and estimator prediction p_i, the estimator receives a calibration penalty via the Brier score:
$$r_i^{\text{est}} = -(y_i - p_i)^2.$$
The estimator's total loss aggregates these penalties. In practice, the estimator is trained purely via supervised calibration loss rather than RL.
- **Prover reward.** The prover receives rewards derived from oracle feedback on the debate trajectory. Rewards encourage the prover to surface subclaims where the estimator is miscalibrated and to produce decompositions leading to correct final answers. During GRPO updates, the prover optimizes expected cumulative episode reward with KL regularization toward its base policy.

Evaluation Metrics

We evaluate both agents (the prover and the estimator) using metrics directly recorded during training. These metrics quantify predictive calibration, policy improvement, stability of optimization, and overall behavioral trends.

In evaluation, we compare the trained prover-estimator model to an untrained one to test if the theoretical guarantees from Brown-Cohen et al. (2025) (Brown-Cohen, Irving, and Piliouras 2025) hold, i.e. if the estimator is calibrated, then the prover will be unable to obfuscate.

Estimator performance. For every training iteration, the estimator's predictions on oracle-labeled subclaims are evaluated using Brier score, which measures the mean squared error between estimator probabilities and oracle truth labels. We additionally track *rolling-average estimator accuracy* (window size 5), which provides a smoothed view of estimator improvement over training .

Prover performance (RL) The prover's GRPO optimization produces the following diagnostics:

- Prover average reward per episode:obtained from oracle-derived trajectory rewards
- KL divergence from reference policy: a measure of policy drift from the base LLM, used for KL-regularized updates

We also compute a rolling-average trend for prover rewards indicating coarse improvement or collapse over time.

Although final-answer accuracy on held-out QuALITY questions is computed separately during evaluation, the training-time metrics reported above serve as important leading indicators of debate quality. In particular, trends in estimator calibration and prover rewards provide early signals about whether the debate dynamics are stabilizing or degenerating over training. When these metrics improve, debates are more likely to converge toward truthful resolutions rather than superficial or obfuscated ones, whereas persistent noise or collapse in these signals often precedes poor generalization at evaluation time.

Results

Our primary goal was to determine whether the prover-estimator debate protocol, instantiated with small open-source models and trained via GRPO, exhibits the qualitative learning dynamics suggested by theory: improving estimator calibration, increasing prover reward, and ultimately converging toward more reliable debate trajectories. In this subsection, we first describe the behavior of a non-learning baseline, then contrast it with the full RL-trained system.

Non–RL baseline (no model updates). Figure 1 presents evaluation metrics for the estimator in a setting where neither the prover nor the estimator is updated between iterations, serving as a static baseline for comparison. We repeatedly run debates on 300 held-out questions and track the average Brier score for each set of three debates, allowing us to assess calibration behavior over time in the absence of learning. While a few episodes appear to achieve a Brier score of zero, closer inspection reveals that these cases arise from failures in rule adherence, where the estimator does not output a valid probability. In such instances, no meaningful score is computed, artificially deflating the reported average rather than reflecting perfect calibration.

All the metrics vary widely across iterations, with no clear sign of convergence. This behavior is expected in a purely sampling-based baseline, where stochasticity in the debate trajectories, the oracle's leaf evaluations, and the estimator's initial beliefs all contribute to substantial variance across episodes. In the absence of parameter updates, there is no mechanism for reducing this variance or improving calibration over time, so fluctuations persist throughout training. Importantly, this baseline provides a reference point for interpreting estimator metrics when *no learning occurs*, allowing us to distinguish genuine learning dynamics from noise introduced by stochastic debate outcomes.

Prover–Estimator behavior under RL When we enable learning and train the estimator as a follower via supervised calibration updates, the resulting training-time metrics are shown in Figure 2. We observe hints of mild improvement early in training, as the Brier score avoids some of the highest spikes and generally lies within a narrower range than in the non-learning baseline. This suggests that the estimator may be partially adapting to the distribution of subclaims it encounters during debate. However, these effects are limited, and the metrics continue to exhibit substantial per-iteration noise throughout training. In particular, we don't observe steady or monotonic improvement typically associated with a well-behaved supervised learner, indicating that the estimator's learning signal remains weak or unstable.

Figure 3 focuses on PPO diagnostics for the prover: average reward and KL divergence from the reference policy. The KL divergence remains almost perfectly flat across training steps, indicating that the GRPO updates induce very little drift away from the base policy. The consistently negative average reward, combined with a flat KL divergence, suggests that the trained prover behaves similarly to a static policy that occasionally "wins" by chance. So replacing the learning prover with a frozen one would likely yield comparable behavior under the same data-collection process.

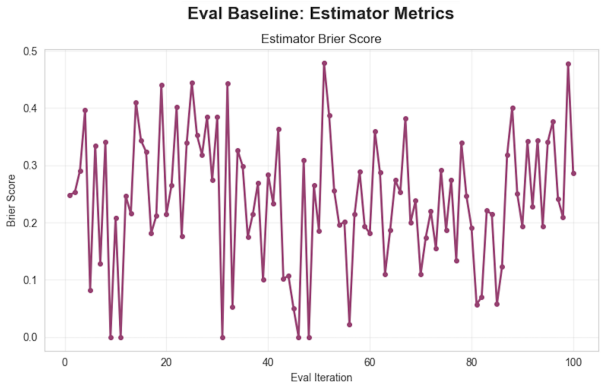

Figure 1: Estimator evaluation metrics in a *non-learning* baseline where prover and estimator parameters are frozen and only debates are sampled. Each point corresponds to an evaluation iteration with no intervening model updates.

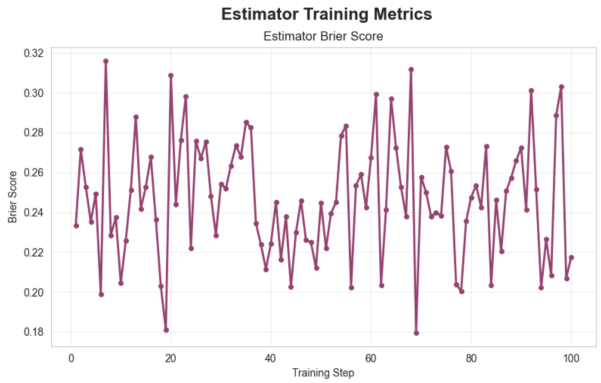

Figure 2: Estimator training metrics when the estimator is updated at every iteration using Brier-score supervision on oracle-labeled subclaims.

To better visualize slow trends, Figure 4 plots rolling averages (window size 5) for estimator accuracy and prover average reward. The smoothed estimator accuracy curve shows minor rises and falls but no clear upward trajectory; it oscillates in a narrow band around approximately 0.6. The prover's rolling-average reward remains mostly negative, with occasional excursions toward zero and a few positive spikes that are not sustained.

We also experimented with training larger base models, specifically Llama-3.2-3b-Instruct (Figure 5) and Llama-3.1-8b-Instruct (Figure 6). Both models were trained on an NVIDIA A100 PCIe GPU. Notably, though, the 8B model used 200GB in VRAM and took four days to train for just 120 iterations, demonstrating the significant computational barriers that must be overcome to enable significant RL findings. For both models, even after training for more iterations than the 1B model, we observe that the estimator's Brier score still heavily fluctuates, and the prover's average reward hovers near 0, which are indicative of the other limitations of this project mentioned in Section 6.2.

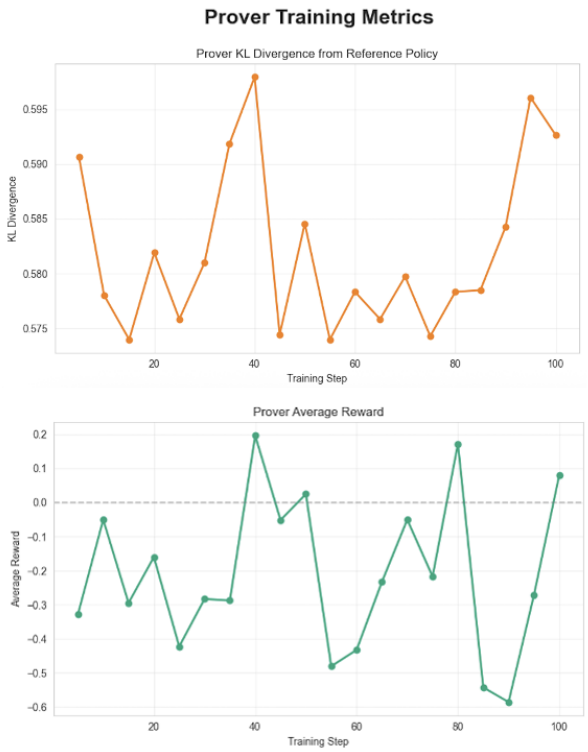

Figure 3: Prover training metrics under GRPO/PPO. Top row: prover average reward, GRPO loss. Bottom row: policy-gradient loss, KL divergence from the reference (base) policy.

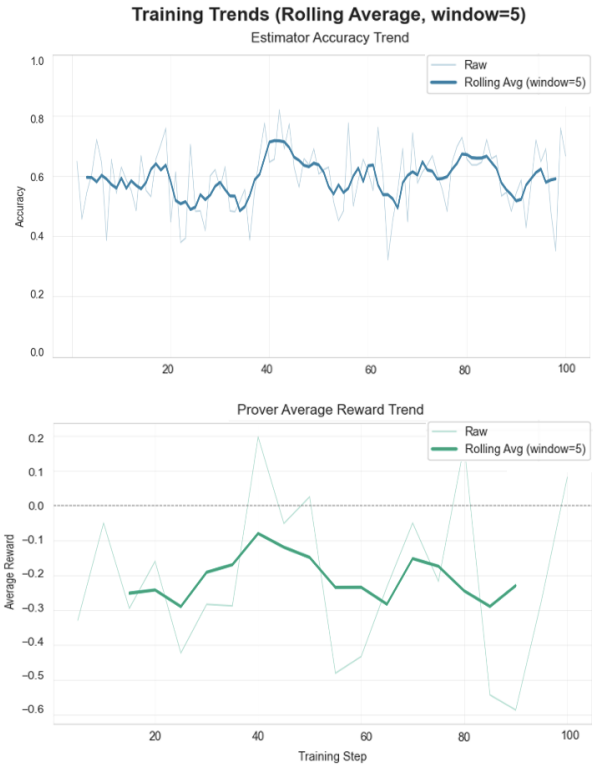

Figure 4: Rolling-average training trends (window size 5) for estimator accuracy (top) and prover average reward (bottom). The raw curves (light lines) are overlaid with smoothed trends (dark lines).

Finally, we include a sample transcript from a debate episode after training the 3B model in Figure 1. In the transcript, we can make two main observations:

1. The prover often struggles to make coherent decompositions. In this example, the first subclaim at $k = 1$ isn't representative of the top-level claim, and the leaf subclaim at $k = 2$ isn't directly related to the first subclaim. In a proper debate structure, we should expect these subclaims to be logically related so that the recursive structure makes sense.

2. The estimator finds it difficult to correctly guess how accurate a subclaim is. In this example, at all levels of the tree, the estimator's guess is near 0.5, a sign that it is extremely uncertain about the truthfulness of the subclaims, even though it is given the original passage in its prompt. This suggests the need for more advanced models, particularly in settings that require long-context reasoning.

These observations further support the limitations we address in Section 6.2 and the future work that we discuss in Section 6.3.

Discussion

Incentive-Optimization Mismatch

We believe that the negative empirical results should not be interpreted as a refutation of the prover-estimator debate framework. Rather, they expose a systematic mismatch between the incentive structure assumed by the theoretical framework and the optimization dynamics induced by gradient-based learning in language models. Table 1 maps the core theoretical assumptions underlying prover–estimator debate to their empirical counterparts and summarizes how each assumption breaks down in our experiments. This mapping clarifies that the observed failures are structural rather than artifacts of scale or implementation fidelity.

Recall the prover–estimator protocol relies on several strong assumptions:

1. the estimator can approximate well-calibrated probabilities over subclaims;
2. the prover can explore and exploit estimator miscalibration by generating structured decompositions; and
3. the oracle reliably evaluates leaf claims so that rewards reflect genuine correctness.

In our setting, all three assumptions are strained. First, the estimator's calibration fails to improve reliably, exhibiting

k	p_k	A_k	Target	Selected Claim (simplified)
2	0.62	+1	MD = 0	Top-level: They will not have to transport more marocca because the plants grow extraordinarily fast and they reproduce on a large-scale
1	0.54	+1	MD = 0	Subclaim: They have already transported a significant amount of marocca to Mypore II.
0	0.54	-1	Oracle = 0	Leaf: The plants have been in Mypore II for a sufficient amount of time to exhibit extraordinary growth.

Table 1: Sample debate transcript from the trained Llama-3.3-3b-Instruct prover.

Theoretical Assumption	**Empirical Proxy**	**Observed Outcome**
Estimator calibration	Training on Brier score	High variance with no consistent improvement
Stable decompositions	Semantic coherence across recursion	Subclaims drift from parent claims
Best-response dynamics	Prover reward; KL divergence	Flat KL divergence and near-zero or negative reward
Reliable oracle feedback	Oracle agreement across calls	Noisy rewards weakly correlated with correctness

Table 2: Assumption → Empirical Failure

high variance across training iterations despite direct supervision via Brier score minimization. This suggests that probabilistic belief modeling over free-form natural-language subclaims is itself a difficult learning problem, especially in long-context settings. Second, the prover fails to meaningfully adapt its behavior under RL: KL divergence from the reference policy remains nearly flat, and average rewards remain negative or near zero. This indicates that the reward signal does not induce a learnable gradient toward exploiting estimator miscalibration.

These failures are not independent. Poor decomposition quality amplifies oracle noise, which in turn degrades estimator calibration, further weakening the prover's learning signal. As a result, the formal incentives specified by the protocol are dominated by implicit optimization pressures such as language-model priors, entropy regularization, and reward sparsity.

From a mechanism-design perspective, this suggests that equilibrium guarantees in prover–estimator debate do not automatically translate into learnable equilibria under standard RL fine-tuning. Bridging this gap likely requires introducing additional structure, supervision, or constraints that explicitly align optimization dynamics with the intended incentives.

Limitations

Several limitations of our experimental setup likely contributed to the lack of clear learning:

Model scale and compute. All experiments were run on either an NVIDIA A100 PCIe or an NVIDIA GeForce GTX 1080 GPU with 8 GB VRAM. These devices are not capable of efficiently training large modern 7B+ parameter models with large batch sizes. As a result, both prover and estimator were instantiated as small 1B parameter instruction-tuned models.

Shallow debate trees and restricted environments. To stay within token and latency budgets, we restricted debate depth and the size of decompositions. This precluded many of the theoretically interesting settings where obfuscation arises: deep trees, long chains of dependent subclaims, and highly asymmetric computational costs across branches. In effect, we only explored a narrow slice of the protocol's design space.

Lack of explicit obfuscation metrics. Although our motivation was robustness to obfuscated arguments, we did not construct a dedicated benchmark with controlled "hidden flaw" decompositions, nor did we log metrics explicitly measuring obfuscation. Since we only observed that our setup fails to learn even in non-adversarial settings, it remains to be seen if similar setups at larger scales serve as a better test of obfuscation.

Weak supervision on decomposition quality. Our training loop optimizes estimator calibration and prover reward, but it does not directly supervise the *structure* of the decompositions. The prover is free to generate subclaims that are redundant, poorly grounded, or only loosely related to the original question. This likely amplifies noise in the oracle's evaluations and makes it harder for the estimator to learn stable probability assignments.

Future Work

Future work should focus on aligning learning dynamics with the intended incentives of the mechanism. We propose the following staged roadmap:

Validate the oracle signal, i.e. reduce reward noise first. Construct a small set of leaf-claim templates with deterministic verification (e.g., span-extraction evidence, exact-match entailment, or synthetic tasks with symbolic ground truth). Measure oracle agreement and reward variance, and only proceed once oracle labels are stable enough to support learning.

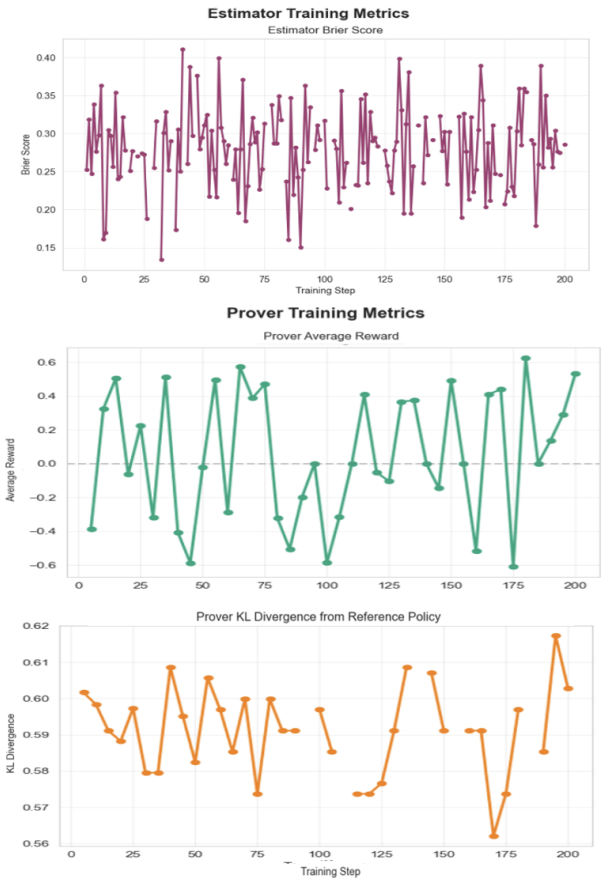

Figure 5: Prover and estimator training metrics using Llama-3.2-3b-Instruct as the base model for both prover and estimator.

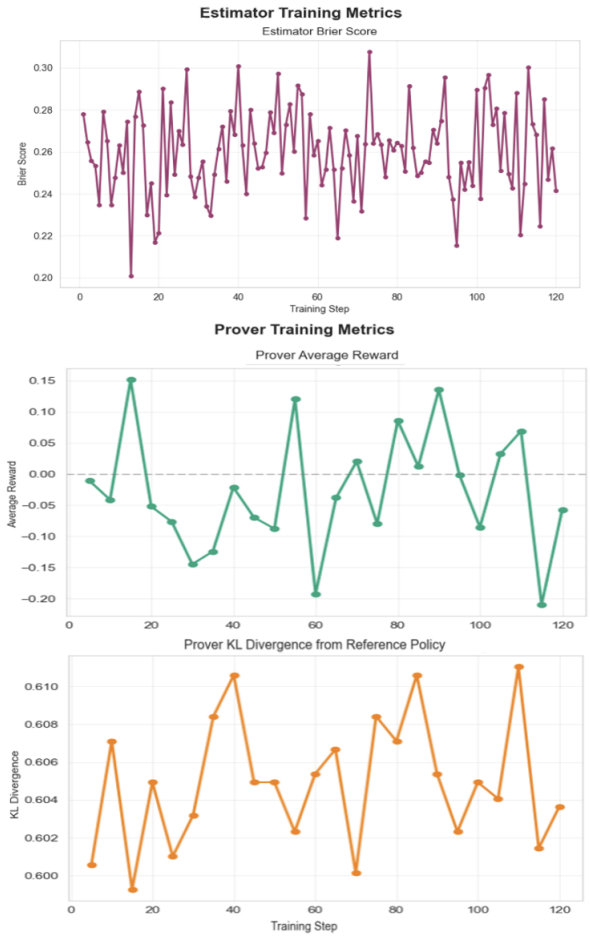

Figure 6: Prover–estimator training metrics with Llama-3.2-8B.

Establish a decomposition-quality baseline before RL Freeze the base model and collect prover-generated decompositions under prompting only. Quantify decomposition coherence (e.g., embedding similarity or NLI entailment between parent and subclaim) and filter/curate a "clean" subset. This produces a measurable starting point for stability and prevents RL from operating on unstructured, drifting subclaims.

Train the estimator offline to a calibration target. Using the curated subclaim dataset, train the estimator purely with supervised calibration loss until Brier score improves and variance narrows on held-out subclaims. This creates a stationary opponent and ensures the estimator meets its intended role before strategic interaction.

Run the estimator-ceiling ablation with a frozen estimator. Replace free-form subclaims with a restricted schema (e.g., claim types such as "quote evidence," "entails," "contradiction," "sub-answer," with required fields). Compare coherence and learning stability to the unconstrained setting. This directly tests whether structure improves the learnability of the intended incentives.

Re-enable joint training with controlled co-adaptation. Move from frozen-estimator training to alternating updates with conservative schedules (e.g., many estimator steps per prover step, slower prover learning rate, stricter KL penalties). Evaluate whether co-adaptation can be stabilized once calibration and decomposition structure are improved.

Stress-test obfuscation explicitly. Only after Steps 1–6 yield stable learning dynamics, evaluate on benchmarks designed for obfuscation: deep trees, hidden flaws, and asymmetric verification costs. Report targeted metrics (e.g., ability to surface hidden errors, robustness to adversarial decompositions), rather than only end-task accuracy.

Conclusion

We present the first empirical instantiation of prover-estimator debate with RL language-model agents, but learning failed to produce calibrated estimators or adaptive provers. This reveals that incentives do not translate into learnable equilibria under gradient-based optimization. Debate-based oversight must account for these dynamics.

References

Amodei, D.; Olah, C.; Steinhardt, J.; Christiano, P.; Schulman, J.; and Mané, D. 2016. Concrete problems in AI safety. *arXiv preprint arXiv:1606.06565*.

Barnes, B. 2020. Debate update: Obfuscated arguments problem. https://www.alignmentforum.org/posts/PJLABqQ962hZEqhdB/debate-update-obfuscated-arguments-problem.

Bi, Z.; Lu, M.; Li, Y.; Roy, S.; Guan, W.; Ziyadi, M.; and Wang, X. 2025. OPTAGENT: Optimizing Multi-Agent LLM Interactions Through Verbal Reinforcement Learning for Enhanced Reasoning. *arXiv preprint arXiv:2510.18032*.

Brown-Cohen, J.; Irving, G.; and Piliouras, G. 2025. Avoiding Obfuscation with Prover-Estimator Debate. arXiv:2506.13609.

Buhl, M. D.; Pfau, J.; Hilton, B.; and Irving, G. 2025. An alignment safety case sketch based on debate. arXiv:2505.03989.

Christiano, P.; Shlegeris, B.; and Amodei, D. 2018. Supervising strong learners by amplifying weak experts. *arXiv preprint arXiv:1810.08575*.

Engels, J.; Baek, D. D.; Kantamneni, S.; and Tegmark, M. 2025. Scaling Laws for Scalable Oversight. *arXiv preprint arXiv:2504.18530*.

Irving, G.; Christiano, P.; and Amodei, D. 2018. AI safety via debate. arXiv:1805.00899.

Khan, A.; Michael, J.; Madaan, A.; and Perez, E. 2021. Debate as supervision for free-form question answering. *NeurIPS 2021 Workshop on Foundation Models*.

Leike, J.; Krueger, D.; Everitt, T.; Martic, M.; Maini, V.; and Legg, S. 2018. Scalable agent alignment via reward modeling: a research direction. *arXiv preprint arXiv:1811.07871*.

Michael, J.; Shlegeris, B.; and Perez, E. 2023. Debate helps supervise unreliable experts. *arXiv preprint arXiv:2311.08702*.

Perez, E.; Ringer, S.; Lukóšiūtė, K.; Nguyen, K.; Chen, E.; Heiner, S.; Pettit, C.; Olsson, C.; Kundu, S.; Kadavath, S.; et al. 2019. Finding cruxes for disagreements with language models. *arXiv preprint arXiv:1911.12237*.