How to Have a Sensitive Debate: An Instance-Optimal Protocol for AI Debate

Jiawei Li*
UT Austin
davidlee@cs.utexas.edu

Zhiyang Xun*
UT Austin
zxun@cs.utexas.edu

Lijie Chen
UC Berkeley
lijiechen@berkeley.edu

Jonah Brown-Cohen
Google DeepMind
jonahbc@google.com

Abstract

As powerful AI systems reach and sometimes surpass the abilities of human experts across a range of cognitively demanding tasks, the problem of accurate oversight and supervision of these systems has become increasingly urgent. One promising approach is AI debate, which seeks to leverage a debate between two powerful AIs to break complex questions down into simpler claims that can be easily judged directly. Theoretical work on debate has formalized this intuition in the language of computational complexity theory, where the goal is to design protocols (i.e., rules of the debate game) that provide rigorous guarantees on correctness for judging solutions to complex problems with limited supervision. Specifically, the current best protocol has been shown to work for all problems that have sufficiently stable decompositions into subproblems. In this paper, we design a new protocol for this same class of problems that improves on the prior work in several ways. First, correctness holds in a worst-case rather than an average-case sense. Second, being honest and correct is a dominant-strategy equilibrium for both debaters, rather than a Stackelberg equilibrium. Finally, we prove black-box lower bounds, showing that our new protocol is instance-wise optimal. That is, no protocol for this class of problems can outperform ours while making only black-box queries to human judgments. We obtain these results by relating the notion of stable problem decompositions to the concept of fractional block sensitivity from query complexity.

1 Introduction

Modern AI systems are trained to perform complex reasoning tasks including autonomous software engineering [Anthropic, 2026], novel mathematical research [Abouzaid et al., 2026], and general economically valuable tasks across a wide range of occupations [Patwardhan et al., 2025]. One of the core drivers of this progress is the ability to design training environments with verifiable reward signals, e.g., numerical solutions or automatically verifiable proofs for mathematics or automated test cases for code. As the scope of tasks performed by AI systems broadens and increases in complexity, it becomes increasingly difficult to provide automatically verifiable rewards that capture all relevant aspects of the task. Furthermore, many tasks of interest have inherently subjective or fuzzy descriptions, so that human verification of AI outputs is necessary. These difficulties lead to the problem of *scalable oversight* — the design of training protocols that leverage the power of rapidly

*Equal contribution. Listing order is random.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

improving AI systems in order to augment human abilities to provide accurate training supervision to these same systems [Leike et al., 2018].

As a motivating example, consider the case of an autonomous AI agent tasked with completing a research project in the quantitative social sciences. The description of the task will likely be high-level and somewhat open-ended, possibly including relevant data sources and the question the analysis is trying to answer. However, in any such project there will be many individual decisions with multiple plausible choices, along with requirements that are not explicitly mentioned, e.g., don't calculate hundreds of different subpopulation statistics from the data, and then choose to analyze the most surprising one post hoc. Hence, directly evaluating the full output of an autonomous AI agent would both require human judgment to some degree and be very difficult even for an expert human to judge directly. Furthermore, the cost of failing to provide accurate supervision in such settings can be quite high. For example, even in slightly misconfigured training environments for software engineering, AI systems often find cheating solutions, such as editing test cases to always pass, rather than actually solving the problem. Worse still, in some such cases AI systems have learned to generalize this negative behavior, resulting in dishonest and harmful actions across many other settings unrelated to coding [MacDiarmid et al., 2025].

One of the primary approaches to addressing the scalable oversight question is *AI debate*, where two powerful AI systems argue about the correctness of an AI output, and a human judges the outcome [Irving et al., 2018]. Intuitively, the idea is that each debater will be able to recursively zoom in on individual flaws in specific pieces of their opponent's arguments, until the original problem has been broken down to a level that is simple enough for a human to judge directly. Critically, there are complexity-theoretic formalizations of this setup that can allow for rigorous definitions and evaluations of what it means for limited human judges to verify the outputs of AI systems that are much too complex to judge.

Prior theoretical research has focused on the design of provably-correct debate protocols, i.e., rules of the game that constrain the AI debaters and specify how the human judge should evaluate their arguments. This prior work encountered a major obstacle known as the *obfuscated arguments problem*, where one debater can efficiently produce a dishonest argument which cannot be efficiently refuted by their opponent [Barnes, 2020]. While it was shown to be possible to obtain provably-correct protocols for certain limited classes of problems [Brown-Cohen et al., 2024], the initial promise of debate for recursively decomposing highly complex problems remained in doubt. Subsequent progress on the obfuscated argument problem was made via the prover-estimator debate protocol of Brown-Cohen et al. [2025], which avoids obfuscated arguments for a class of problems that admit *sufficiently stable* decompositions into subproblems. Despite this progress on expanding the class of problems that can be handled by debate, the prover-estimator protocol has additional limitations when compared to prior work: it provides average-case guarantees that hold for a weaker equilibrium concept than previously studied.

In this paper we design a new debate protocol that provides worst-case guarantees for the original, stronger dominant-strategy equilibrium concept. The protocol handles exactly the same class of sufficiently stable problems as prover-estimator debate, while also being simpler to execute and analyze. We further prove that this protocol is instance optimal unless further structural assumptions are made on the human judgments used to verify debates.

1.1 Our Results

Our new protocol is simple to describe. The first debater, Alice, decomposes the original claim into sub-claims. The second debater, Bob, then has two choices: either (1) demonstrate that Alice's decomposition is too *sensitive* to small errors in the sub-claims, or (2) choose a sub-claim on which to continue the debate recursively. The key to the analysis is the equivalence between the notion of stable decompositions required for prover-estimator debate and bounds on *fractional block sensitivity*, a quantity defined in the query complexity literature which is tightly connected to the number of queries a randomized verifier must make to check an untrusted proof [Kulkarni and Tal, 2016]. In particular, an upper bound on the fractional block sensitivity of Alice's decomposition implies an efficiently computable strategy for Bob to select a sub-claim that has a non-trivial chance of being incorrect.

The equivalence to fractional block sensitivity further enables us to prove that the performance of our protocol is *instance optimal* assuming black-box access to human judgments. That is, if the particular problem instance admits a decomposition with fractional block sensitivity ρ, then *any protocol* which correctly decides the outcome on this instance cannot have asymptotically lower complexity for the debaters. All prior theoretical work on debate has modeled human judgment as a black-box oracle, so this result can be interpreted as demonstrating the necessity of going beyond this model in order to further expand the class of problems solvable via debate.

1.2 Technical Overview

We next give an informal mathematical overview of our protocol and proof techniques. We will begin by describing how the protocol works in our running example of an autonomous AI agent Alice performing quantitative social science research.

The protocol in action. In order to efficiently evaluate the correctness of the final research paper and codebase produced by the quantitative social science agent, we use a debate between Alice, who defends her work, and another AI Bob who attempts to find flaws. To determine the validity of a claim in the paper, we ask Alice to decompose the original claim x into sub-claims $y_1, \ldots, y_q$. One can view these sub-claims as pieces of supporting evidence. For example, the claim x might be that a certain statistical modeling assumption is justified. Suppose that we would believe this claim if a majority of correct, published papers in the existing literature supported it. In this case, the sub-claims y_i are of the form: this particular paper from the existing literature is both correct and supports the original claim. Furthermore, the truth of the original claim x is the majority vote of the truth values of the sub-claims y_i.

Next, suppose that Alice claims that in fact $2/3$ of the published papers y_i support the claim. We view this as an assignment by Alice of alleged truth values z_i to each sub-claim y_i. In this case, the key insight is that Bob need not check every single paper for correctness in order to verify the overall claim. Instead, Bob can just randomly sample a paper among the $2/3$ that Alice says support the original claim, i.e., those where $z_i = 1$. If Alice is lying about the original claim x, then at least $1/6$ of the sub-claims must also be lies. Hence, by sampling randomly among the $2/3$ where $z_i = 1$, Bob will have a $\frac{1/6}{2/3} = 1/4$ chance of catching Alice in a lie. Thus, just by sampling a few research papers from the entire area to check, Bob can be confident of catching Alice in a lie.

On the other hand, suppose that Alice claims that 50 out of a total of 99 papers y_i support the original claim x. In this case, Bob has no recourse but to check every single one of the 50 papers, as even a single incorrect paper would flip the majority vote and overturn the truth of the original claim. The key to our protocol is that Bob is allowed to demonstrate that Alice has made an unreasonably sensitive claim of this form, in which case Alice will lose the debate. This in turn requires Alice to only decompose her claims in a way that is not too sensitive to errors in individual sub-claims. We view this as a very natural restriction when arguing about claims that inherently involve fuzzy human judgment. Indeed, one should be very suspicious of any argument that is highly sensitive to changes in individual pieces of evidence, and yet relies on fuzzy human judgments in each step.

Formalization via fractional block sensitivity. Our protocol follows the same outline as described in the example: Alice decomposes a claim x into sub-claims y_i, Bob either objects that Alice's decomposition is too sensitive, or instead selects a sub-claim to debate further. The key technical component is the notion of *fractional block sensitivity* which generalizes the intuition above about the majority function to all possible aggregation functions Alice might choose. In particular, the fractional block sensitivity ρ is defined as the optimum of a linear program that depends on the aggregation function and Alice's alleged truth values z_i for the sub-claims y_i. In the case of the majority vote function with $2/3$ of the truth values z_i set to 1, the optimum of the linear program is $\rho = 4$, which corresponds to Bob's chance of $1/4$ of catching Alice in a lie if he samples a random y_i among the $2/3$ claimed true. This correspondence is not a coincidence–it generalizes to any aggregation function. More formally, the dual solution of the fractional block sensitivity linear program directly yields probabilities p_i for Bob to sample sub-claims. When sampling sub-claims with these probabilities, Bob has at least a chance of $1/\rho$ of catching Alice in a lie. See Section 3.3 for the precise formal definition.

The protocol runs recursively on the sub-claim sampled by Bob, until it reaches a depth of d, whereupon a human judge directly evaluates the final selected sub-claim. Unless Alice wants to lose immediately, she must ensure that in each round the decomposition into sub-claims that she makes has fractional block sensitivity bounded by ρ. Thus, we conclude that if Bob follows the sampling scheme provided above, he will catch Alice in a lie with probability at least ρ^{-d}. Hence, by running the protocol $O(\rho^d \log n)$ times, Bob will catch Alice in a lie with probability at least $1 - O(1/\operatorname{poly}(n))$. See Section 4 for details. To make sense of this result, recall that Alice decomposes each original claim into q sub-claims. So exploring the full tree of sub-claims would require time q^d. As shown in the example of the majority vote aggregation function, one can have $q = 99$, while $\rho = 4$, implying that ρ^d can be significantly smaller. In the standard setting, the debate has depth $d = O(\log n)$, implying that our protocol allows for a polynomial in n speed-up beyond exhaustive search of Alice's decomposition.

Honesty and correctness are dominant strategies. Our protocol for fractional block sensitivity (FBS) debate satisfies the key properties of *completeness* and *soundness*. Completeness requires that if the original claim x is correct, then no matter what strategy Bob employs, Alice can always win by honestly and correctly decomposing every claim. Similarly, soundness requires that if the original claim x is incorrect, then no matter what strategy Alice employs, Bob can win with overwhelming probability by honestly following the prescribed protocol. In other words, honestly and correctly following the prescribed protocol is a *dominant strategy* for each player in the debate game. This is in contrast to the prior prover-estimator debate protocol, which only guarantees soundness under the following Stackelberg equilibrium condition: for any dishonest strategy selected by Alice, there exists a winning strategy by Bob, which in general depends on Alice's strategy in a potentially complicated way. This is a technically much weaker condition, and in practice training for dominant strategy equilibria is much more stable and tractable than Stackelberg equilibria.

Furthermore, our protocol applies on a worst-case basis: if an individual claim x is incorrect, our protocol will allow Bob to efficiently identify a flawed sub-claim for the human judge. In contrast, the prover-estimator protocol only satisfies soundness on average over a distribution on inputs. This is sufficient to guarantee that training the debaters to maximize their expected win rates leads to a debate equilibrium where the debaters are honest on all but a small fraction of questions. However, our protocol for FBS debate goes further by guaranteeing that on each individual question, honesty is the best policy.

Instance-wise optimality. Finally, we show in Section 5 that any protocol with only black-box access to human judgments must make at least $\Omega(\rho^d)$ queries to find a flawed sub-claim with constant probability, matching the cost of our FBS debate protocol. This lower bound holds instance-wise: if the best decomposition of an original claim x into a tree of sub-claims has fractional block sensitivity ρ in each step, then ρ^d queries are required on that particular instance x. The assumption of black-box access to human judgments is critical for the lower bound to hold. One can view this instance-wise optimality as demonstrating that the black-box judgment assumption, which is common to all prior theoretical work on debate, must be relaxed in order to make further progress on debate protocols.

2 Related work

Prior work on the complexity-theoretic formulation of debate began with Irving et al. [2018], whose protocol works for computationally unbounded provers. Subsequently, Barnes [2020] pointed out the obfuscated arguments problem, demonstrating that current protocols would fail in the case that the AI debaters were equally computationally powerful, but still computationally bounded. Brown-Cohen et al. [2024] introduced doubly-efficient debate, which works for computationally bounded debaters, but solves a more limited class of problems. Most closely related to our work is that of Brown-Cohen et al. [2025], which introduced the prover-estimator debate protocol to handle a broader class of recursively decomposable problems for computationally bounded debaters. Prover-estimator debate introduced the notion of stability (equivalent to fractional block sensitivity) that we study in this paper.

The idea of having computationally unbounded provers convince a polynomial-time verifier was already present in the work on alternating Turing machines of Chandra et al. [1981]. Another complexity-theoretic model of arguments involving private communication between adversarial

provers and a judge is due to Feige and Kilian [1997]. Finally, competing provers have been studied in the context of secure delegation of computation by Canetti et al. [2013]. On the algorithmic side, Chen et al. [2023] designed algorithms for the problem of learning the equilibrium in the zero-sum game defined by a debate protocol. Here the focus is on ensuring the learning algorithms and the effort of the judge are computationally efficient while the strategy space may be exponentially large.

There have also been several recent works on utilizing more classical single-prover interactive proofs in machine learning. This has been done most frequently with the goal of verifying that the output of an AI is correct [Amit et al., 2024, Anil et al., 2021, Hammond and Adam-Day, 2025, Chen et al., 2026], while some research has also focused on using these techniques for interpretability and legibility [Wäldchen et al., 2022, Kirchner et al., 2024]. In particular, Chen et al. [2026] construct doubly-efficient interactive proofs and arguments for *oracle* circuits when either the computation is *robust* to corruption of a small fraction of its oracle transcript or the oracle has low-degree algebraic structure. Their robustness condition is a uniform-weight special case of bounded FBS at the level of a fixed one-step aggregator, while our local FBS condition permits an optimized nonuniform sampling distribution.[1] Their full framework remains incomparable to ours: their protocol assumes an explicit circuit fixed at the beginning for the target computation, whereas ours uses an (online) recursive decomposition framework with a fixed one-step aggregator. Their low-degree-oracle result is orthogonal to our lower bound because it imposes additional algebraic structure on the oracle.

Several empirical works have measured the performance of actual LLMs in setups inspired by theoretical debate protocols [Michael et al., 2023, Kenton et al., 2024, Khan et al., 2024]. These results provided some evidence that debate helps certain limited LLM judges more accurately verify the output of other LLMs. Finally, large-scale training for mathematical reasoning has relied on the power of a debate-like training protocol in which one model critiques a proof and another model further critiques the critique [Shao et al., 2025].

3 Preliminaries

3.1 Complexity-theoretic framework of AI debate

We adopt a complexity-theoretic framework (cf. Brown-Cohen et al. [2025]) for studying the power of AI debate. We represent the human judgment as an oracle $\mathcal{O} : \{0,1\}^n \to \{0,1\}$, where the value of $\mathcal{O}(u) \in \{0,1\}$ corresponds to the human judgment on a particular question u. Both debaters and the judge are algorithms modeled as Turing machines given black-box query access to the oracle $\mathcal{O}$. We denote by $M^{\mathcal{O}}(x)$ the result of running algorithm M on input x given access to $\mathcal{O}$.

The aim of debate is to design protocols that solve broad classes of computational problems while using few queries to the human-judgment oracle $\mathcal{O}$. Any decision problem (yes-or-no problem) can be represented by a *language* $L \subseteq \{0,1\}^*$, where $x \in L$ means that the correct answer on input x is "yes." Equivalently, we may view L as a Boolean function $L : \{0,1\}^n \to \{0,1\}$, with $L(x) = 1$ iff $x \in L$.

3.2 Recursively decomposable problems

In this paper, we focus on the class of problems L that exhibit a recursive decomposition. There are two components of a recursive decomposition: a *one-step aggregator*, and a *decomposition tree*. Intuitively, if the original problem x is decomposed into q sub-problems $y = (y_1, \ldots, y_q)$ in one step, then a one-step aggregator should return the correct answer for x when given the correct answers to y.

Definition 3.1 (One-step aggregator)**.** Let L be a language. A width q one-step aggregator M_D for L is an algorithm that takes as input a problem instance $x \in \{0,1\}^n$, a sequence of q *sub-problems* $y = (y_1, \ldots, y_q) \in \{0,1\}^{n \times q}$, and a sequence of q *answers* $z = (z_1, \ldots, z_q) \in \{0,1\}^q$. The pair (x, y) is called a one-step decomposition of x.

The one-step aggregator M_D should either accept or reject a one-step decomposition (x, y). If M_D rejects a one-step decomposition (x, y), then $M_D(x, y, z) = \perp$ for any z; otherwise, $M_D(x, y, z) \in \{0,1\}$ for any z. Furthermore, if M_D accepts a one-step decomposition (x, y), then $M_D(x, y, z) = L(x)$ if z gives the correct answers to y, i.e., $z_i = L(y_i)$ for all $i \in [q]$.

[1] Their ε-robustness assumption implies $\mathrm{fbs} = O(1/\varepsilon)$ for the induced Boolean function: uniformly assigning weight $1/(\varepsilon\ell)$ to each of the ℓ oracle-answer coordinates gives a feasible solution to the dual FBS program.

In the setting of AI debate, we assume the one-step aggregator M_D is a rather simple and efficient algorithm that all parties agree on at the beginning of the protocol. One simple example of M_D is taking the AND of all answers z_i, which captures a classic debate strategy — if all the supporting sub-claims are correct, then the original claim is also correct. We will be interested more generally in one-step aggregators that make fuzzier use of supporting sub-claims. To return to the running example of the autonomous agent performing quantitative social science research, the one-step aggregator might take the form: if the majority of published papers in an area support the original claim, then the claim is true. In this case, the supporting sub-claims y_i correspond to the set of published papers in the area, and the aggregator M_D is the majority vote of z_i, each of which represents whether or not the paper y_i supports the original claim.

We further require that the property of M_D as defined in Definition 3.1 can be easily verified by human judgment. Note that in this definition, M_D is allowed to reject a possibly correct but hard-to-verify one-step decomposition; it could still be useful as long as it accepts most of the obviously valid decompositions. Given this, we can assume that M_D never outputs $\perp$ in a debate protocol without loss of generality: If any debater proposes a one-step decomposition rejected by M_D, then we can make that debater immediately lose the debate, so any clever debater would avoid that.

We now define the decomposition tree using the one-step aggregator.

Definition 3.2 (Decomposition Tree)**.** Let L be a language and fix a width q one-step aggregator M_D for L. A depth d, width q *decomposition tree* $\mathcal{T}$ for input $x \in \{0,1\}^n$ is defined as follows.

1. Each tree node u corresponds to a problem instance $\hat{x}_u \in \{0,1\}^n$ for language L; the root node at level $k = d$ corresponds to the input x.
2. For each non-leaf node u at level $k > 0$, we have $L(\hat{x}_u) = M_D(\hat{x}_u, y_u, z_u)$, where $y_u = (y_{u,1}, \ldots, y_{u,q})$ are the sub-problems from the child nodes of u at level $k-1$, and $z_u = (L(y_{u,1}), \ldots, L(y_{u,q}))$.
3. For each leaf node u at level $k = 0$, there exists a single query q_u such that $L(\hat{x}_u) = \mathcal{O}(q_u)$.

Unlike the one-step aggregator M_D, which is fixed at the beginning, a decomposition tree $\mathcal{T}$ for a given input x is usually generated on the fly as the debate protocol proceeds. A brute-force debate protocol that visits all the leaves of a decomposition tree requires q^d queries to the oracle $\mathcal{O}$.[2] A better debate protocol based on recursive decomposition should avoid checking most of the nodes in the decomposition tree.

3.3 Fractional block sensitivity

The main structural assumption that characterizes the class of problems solvable by our protocol is that the functions arising in the decomposition of the problem have bounded *fractional block sensitivity*. Fractional block sensitivity is a metric from query complexity defined for a Boolean function f. Intuitively, it captures how closely a distribution can match its individual marginals to a fixed input x while still flipping the output of f.

To gain an intuition for this notion, consider the majority vote function $f(x) = \textsc{Majority}(x)$ and the string $x \in \{0,1\}^n$ where the first 2/3rds of the inputs are 1 and the rest are 0. In this case, in order to change the output of $f(x)$, one would need to flip at least 1/6th of the input bits among the 2/3rds that are 1. Thus, in any distribution which flips the output, the marginal distribution on each input coordinate in the first 2/3rds must be off by at least $\frac{1/6}{2/3} = 1/4$. In this case the fractional block sensitivity of f would be 4.

On the other hand, for the Boolean AND function $f(x) = \textsc{And}(x)$ on the all-ones input, it is possible to flip the output of f by selecting just one input bit uniformly at random to flip to 0. The marginals of this distribution are all $\frac{1}{n}$, and so the fractional block sensitivity is n. This is a much higher fractional block sensitivity than for the majority vote function, capturing the idea that it is much easier to hide a single mistake that changes the entire outcome for the AND function than it is for the majority function. In general, fractional block sensitivity generalizes this intuition that one should not be able

[2] One can apply the *doubly-efficient* debate protocol Brown-Cohen et al. [2024] here to reduce the number of queries to $\mathcal{O}$ to only 1. However, the time complexity of this protocol is still $O(q^d)$, which is infeasible in most cases.

to stealthily flip the output of a function while maintaining a very similar marginal distribution on input bits.

Definition 3.3 (Fractional block sensitivity)**.** Let $f : \{0,1\}^n \to \{0,1\}$ and $x \in \{0,1\}^n$. For a nonempty set $B \subseteq [n]$, let x^B denote the string obtained from x by flipping the bits in the coordinates of B, i.e.,

$$(x^B)_i \;=\; \begin{cases} 1 - x_i & \text{if } i \in B, \\ x_i & \text{if } i \notin B. \end{cases}$$

A set $B \subseteq [n]$ is a *sensitive block for f at x* if $f(x) \neq f(x^B)$. Write $\mathcal{S}_x(f)$ for the family of all sensitive blocks at x.

The *fractional block sensitivity of f at x* is the optimal value of the linear program

$$\mathrm{fbs}_x(f) \;=\; \max \left\{ \sum_{B \in \mathcal{S}_x(f)} w_B \;\middle|\; \begin{array}{l} w_B \geq 0 \text{ for all } B \in \mathcal{S}_x(f), \\ \displaystyle\sum_{\substack{B \in \mathcal{S}_x(f) \\ i \in B}} w_B \;\leq\; 1 \quad \text{for all } i \in [n] \end{array} \right\}.$$

The *fractional block sensitivity of f* is

$$\mathrm{fbs}(f) \;=\; \max_{x \in \{0,1\}^n} \mathrm{fbs}_x(f).$$

Definition 3.4 (Dual of fractional block sensitivity)**.** Fix $f : \{0,1\}^n \to \{0,1\}$ and $x \in \{0,1\}^n$, and let $\mathcal{S}_x(f)$ denote the family of sensitive blocks at x (as above). The dual linear program to the definition of $\mathrm{fbs}_x(f)$ is

$$\mathrm{fbs}_x(f) \;=\; \min \left\{ \sum_{i=1}^{n} t_i \;\middle|\; \begin{array}{l} t_i \geq 0 \quad \text{for all } i \in [n], \\ \displaystyle\sum_{i \in B} t_i \;\geq\; 1 \quad \text{for all } B \in \mathcal{S}_x(f) \end{array} \right\}.$$

The dual assigns nonnegative weights $\{t_i\}_{i \in [n]}$ to coordinates so that every sensitive block B has total weight at least 1; it minimizes the total weight $\sum_i t_i$. Thus the dual is a fractional *hitting set/cover* of the sensitive blocks by coordinates.[3] By strong duality, its optimum equals the primal value $\mathrm{fbs}_x(f)$.

4 Protocol

In this section, we present our debate protocol based on the fractional block sensitivity of the recursive decomposition. We then prove that our protocol satisfies both the completeness and soundness conditions in the worst case.

FBS Debate Protocol

Setup: Access to the input $x \in \{0,1\}^n$, a language L, the human-judgment oracle $\mathcal{O}$, a one-step aggregator M_D with width q, the depth of the decomposition tree d, and a hyperparameter ρ.

Initialization: Alice outputs a bit $A(x)$ as the answer to $L(x)$.

At level $k \in \{d, d-1, \ldots, 1\}$:

1. Alice decomposes the current problem $\hat{x}$ into q sub-problems $y = (y_1, \ldots, y_q)$, and then outputs a sequence $z = (z_1, \ldots, z_q)$ as the answers to $y_1, \ldots, y_q$. $M_D(\hat{x}, y, z)$ must be equal to the answer Alice claimed for $\hat{x}$ in the previous round, denoted by $A(\hat{x})$; otherwise Alice loses immediately.
2. Bob checks whether $\mathrm{fbs}_z\left(M_D(\hat{x}, y, \cdot)\right) \leq \rho$ by solving the dual LP. If not, Alice immediately loses.
3. If true, let $p_i := t_i / \sum_{j=1}^{q} t_j$, where t is the optimal solution to the dual LP. With probability p_i, Bob will pick the i-th sub-problem y_i for recursion, set $\hat{x} \leftarrow y_i$ and $A(\hat{x}) \leftarrow z_i$, and continue to level $k - 1$.

At the leaf level ($k = 0$): If the answer Alice claimed for the current sub-problem matches $\mathcal{O}$, Alice wins; otherwise Bob wins.

[3]The dual of fractional block sensitivity is essentially the *randomized certificate complexity* defined in Aaronson [2008].

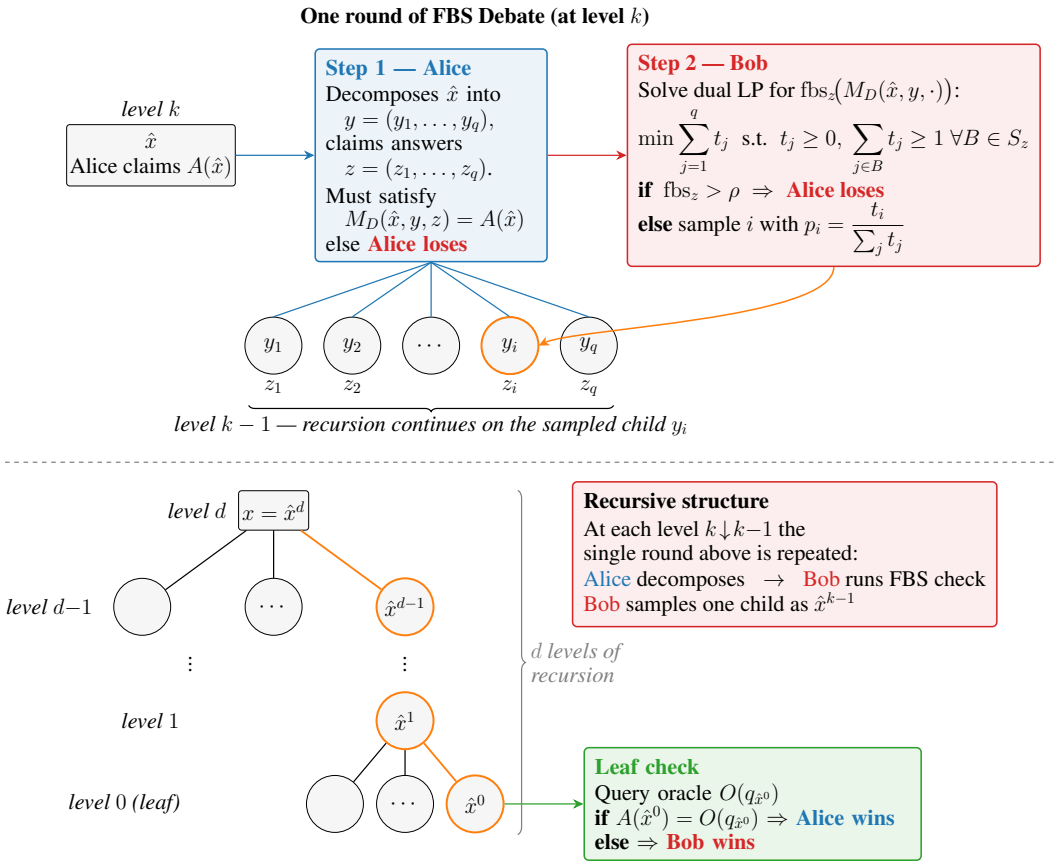

Figure 1: The FBS debate protocol. *Top:* one round at level k. *Bottom:* the recursion for d levels.

We show that an honest Alice can win the debate if Alice can find a decomposition tree solving the problem with low fractional block sensitivity.

Theorem 4.1 (Completeness)**.** *Let T, T_D, Q be three parameters. An honest Alice can always win in the FBS debate on input x in $O\big(d \cdot (T_D + q \cdot T)\big)$ time and $O(d \cdot q \cdot Q)$ queries to $\mathcal{O}$ if there exists a decomposition tree $\mathcal{T}$ for x such that*

- *Alice can calculate the children of any non-leaf node in $\mathcal{T}$ in time T_D.*
- *Alice can solve $L(\hat{x}_u)$ at any node u in $\mathcal{T}$ in time T and Q queries to $\mathcal{O}$;* *(Solvability)*
- *For every non-leaf node u in $\mathcal{T}$, if y are the sub-problems in the child nodes of u and z are the correct answers to y, then $\mathrm{fbs}_z\left(M_D(\hat{x}_u, y, \cdot)\right) \leq \rho$.* *(FBS condition)*

Proof. The honest strategy takes $A(x) = L(x)$, decomposes exactly as $\mathcal{T}$ at every node, and sets z to be the correct answers at every decomposition step. Then (i) the consistency check always passes by the solvability condition; (ii) the stability check always passes due to the FBS condition; and (iii) at depth 0 the oracle agrees with Alice. An induction on depth $k = d, d-1, \ldots, 0$ shows that the run never triggers a losing condition, hence Alice wins with probability 1.

It is easy to check that this strategy takes $O\big(d \cdot (T_D + q \cdot T)\big)$ time and $O(d \cdot q \cdot Q)$ queries to $\mathcal{O}$. □

We then show that whenever Alice answers the top-level problem incorrectly, the protocol will catch her mistake with probability ρ^{-d}, where d is the depth of the recursive decomposition and ρ is the fractional block sensitivity. This immediately implies that by running the protocol $O(\rho^d \log n)$ times, any incorrect answer by Alice will be caught with probability at least $1 - 1/\operatorname{poly}(n)$. Viewed from the lens of practical AI debate, this represents a trade-off between allowing highly sensitive arguments and the inference-time computation required to verify them.

Theorem 4.2. *Suppose $A(x) \neq L(x)$. Then, regardless of Alice's strategy,*

$$\mathbb{P}[\text{Alice loses}] \;\geq\; \rho^{-d}.$$

We refer the reader to Appendix A for the proof of Theorem 4.2.

5 Lower Bound

In this section, we prove an instance-wise lower bound for debate protocols in the recursive decomposition framework. The bound is governed by the fractional block sensitivity of the local aggregation functions.

Setup. Fix a language L, a one-step aggregator M_D, an input x, and a depth-d visible tree $\mathcal{T}$ for x with respect to M_D. For every internal node u, let $y_u = (y_{u,1}, \ldots, y_{u,q})$ be its children and let

$$\phi_u(z) \coloneqq M_D(\hat{x}_u, y_u, z)$$

be its local aggregator. For any oracle $\mathcal{O}$, define the value of a node by bottom-up evaluation:

$$\mathrm{val}_{\mathcal{O}}(u) \coloneqq \begin{cases} \mathcal{O}(q_u) & \text{if } u \text{ is a leaf,} \\ \phi_u\big(\mathrm{val}_{\mathcal{O}}(u_1), \ldots, \mathrm{val}_{\mathcal{O}}(u_q)\big) & \text{if } u \text{ is internal,} \end{cases}$$

where $u_1, \ldots, u_q$ are the children of u. For the lower bound, we treat $\mathcal{T}$ as a syntactic transcript whose truth value under $\mathcal{O}$ is $\mathrm{val}_{\mathcal{O}}(\mathrm{root})$.

We assume that the leaf queries q_ℓ are pairwise distinct. This assumption is needed for a non-trivial lower bound: for example, if all q leaves of a q-ary AND root query the same string, then the decomposition tree $\mathcal{T}$ computes only a one-bit identity function.

Adaptive checker. To strengthen the lower bound, we assume that the second debater Bob is honest and assists the judge, while Alice may use an arbitrary strategy. We model Bob and the judge together as a single *adaptive checker*. Specifically, an m-query adaptive checker $\mathcal{V}$ for $\mathcal{T}$ is a randomized algorithm that receives the full visible tree $\mathcal{T}$, makes at most m adaptive queries to $\mathcal{O}$, may communicate arbitrarily with Alice, and finally outputs a bit intended to equal $L(x)$. Giving the checker an honest Bob and the full visible tree can only make its task easier, so a lower bound in this model also applies to the original debate protocol.

Fix a reference oracle $\mathcal{O}_Y$. For any alternative oracle $\mathcal{O}_N$, let Acc_Y and Acc_N denote the events that $\mathcal{V}$ accepts under $\mathcal{O}_Y$ and $\mathcal{O}_N$, respectively, with Alice using the same fixed strategy in both executions. When the two oracles give different values at the root, $|\,\mathbb{P}[\mathsf{Acc}_Y] - \mathbb{P}[\mathsf{Acc}_N]|$ measures the checker's oracle-distinguishing advantage. We show that if every local aggregator has fractional block sensitivity at least ρ at the child values induced by $\mathcal{O}_Y$, then some alternative oracle changes the root value while giving any m-query checker distinguishing advantage at most $m\rho^{-d}$.

Theorem 5.1. *Let $\rho \geq 1$, and suppose every internal node u satisfies*

$$\mathrm{fbs}_{\boldsymbol{z}^u}(\phi_u) \geq \rho,$$

where $\boldsymbol{z}^u \in \{0,1\}^q$ is the vector of child values induced by $\mathcal{O}_Y$. Then for every m-query adaptive checker $\mathcal{V}$ for $\mathcal{T}$, there is an oracle $\mathcal{O}_N$ such that

$$\mathrm{val}_{\mathcal{O}_N}(\mathrm{root}) \neq \mathrm{val}_{\mathcal{O}_Y}(\mathrm{root})$$

and

$$|\mathbb{P}[\mathsf{Acc}_Y] - \mathbb{P}[\mathsf{Acc}_N]| \leq \min\{1, m\rho^{-d}\}.$$

We refer the reader to Appendix B for the proof of Theorem 5.1.

6 Discussion and Open Problems

Our protocol and matching lower bound show that fractional block sensitivity precisely characterizes the set of problems that can be efficiently solved by debate within this recursive-decomposition framework. Notably, the protocol does not require Bob to determine whether or not Alice's answers to subproblems are correct, but only to understand the structure of the argument well enough to sample subproblems to which the outcome is sufficiently sensitive. This stands in contrast to the original Debate = PSPACE result of Irving et al. [2018], which requires Bob to actually directly determine which of Alice's subproblems are flawed. Of course, this comes with the limitation that the identification of flawed subproblems might actually have much greater computational complexity than creating the subproblems. This is not an issue if we assume that both Alice and Bob have the power to solve any problem in PSPACE, but is a serious obstacle if the debaters are assumed to be computationally bounded (as they will be in practice).

Given this discussion, the question remains: is it possible to design a protocol that goes beyond the fractional block sensitivity barrier we identify? One might hope that utilizing Bob's ability to solve the same problems as Alice could help with this issue, perhaps by allowing Bob to take advantage of Alice's computational limitations in some way. Alternatively, our lower bounds only apply when human judgment is modeled as a black box, accessible only via queries. Classical results in interactive proofs like IP = PSPACE show that opening up the black box yields much stronger protocols. It seems implausible that these methods can be adapted to the case of human judgment—one would need to essentially be able to "read the code" that implements the human judgment. However, it is still possible that some other relaxation of the black-box oracle assumption could lead to better debate protocols.

Broader impacts. This work aims to improve scalable oversight by identifying conditions under which debate can make incorrect claims easier to detect. Its guarantees rely on the stated decomposition, oracle, and debater assumptions; they do not by themselves establish the safety of practical AI debate systems. Applying the protocol when these assumptions fail could create unwarranted confidence in incorrect or harmful outputs. Empirical validation and independent safeguards are therefore necessary before relying on such systems in consequential settings.

Acknowledgments and Disclosure of Funding

JL is supported by Scott Aaronson's Coefficient Giving grant. ZX is supported by the NSF AI Institute for Foundations of Machine Learning (IFML). The authors declare no additional competing interests beyond their listed affiliations.

References

Scott Aaronson. Quantum certificate complexity. *Journal of Computer and System Sciences*, 74(3): 313–322, 2008.

Mohammed Abouzaid, Andrew J. Blumberg, Martin Hairer, Joe Kileel, Tamara G. Kolda, Paul D. Nelson, Daniel Spielman, Nikhil Srivastava, Rachel Ward, Shmuel Weinberger, and Lauren Williams. First proof, 2026. URL https://arxiv.org/abs/2602.05192.

Noga Amit, Shafi Goldwasser, Orr Paradise, and Guy Rothblum. Models that prove their own correctness. *arXiv preprint arXiv:2405.15722*, 2024.

Cem Anil, Guodong Zhang, Yuhuai Wu, and Roger Grosse. Learning to give checkable answers with prover-verifier games. *arXiv preprint arXiv:2108.12099*, 2021.

Anthropic. Claude opus 4.6 system card. System card, Anthropic, February 2026. URL https://www.anthropic.com/claude-opus-4-6-system-card.

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

Jonah Brown-Cohen, Geoffrey Irving, and Georgios Piliouras. Scalable AI safety via doubly-efficient debate. In *Forty-first International Conference on Machine Learning*, 2024.

Jonah Brown-Cohen, Geoffrey Irving, and Georgios Piliouras. Avoiding obfuscation with prover-estimator debate, 2025. URL https://arxiv.org/abs/2506.13609.

Ran Canetti, Ben Riva, and Guy N Rothblum. Refereed delegation of computation. *Information and Computation*, 226:16–36, 2013.

Ashok K. Chandra, Dexter C. Kozen, and Larry J. Stockmeyer. Alternation. *J. ACM*, 28(1): 114–133, January 1981. ISSN 0004-5411. doi: 10.1145/322234.322243. URL https://doi.org/10.1145/322234.322243.

Liyan Chen, Yael Tauman Kalai, and Zoe Xi. How to avoid debate: Scalable ai safety via doubly-efficient interactive proofs. *arXiv preprint arXiv:2607.03561*, 2026.

Xinyi Chen, Angelica Chen, Dean Foster, and Elad Hazan. Playing large games with oracles and AI debate. *arXiv preprint arXiv:2312.04792*, 2023.

Uriel Feige and Joe Kilian. Making games short. In *Proceedings of the twenty-ninth annual ACM symposium on Theory of computing*, pages 506–516, 1997.

Justin Gilmer, Michael Saks, and Srikanth Srinivasan. Composition limits and separating examples for some boolean function complexity measures. *Combinatorica*, 36(3):265–311, June 2016. ISSN 0209-9683. doi: 10.1007/s00493-014-3189-x. URL https://doi.org/10.1007/s00493-014-3189-x.

Lewis Hammond and Sam Adam-Day. Neural interactive proofs. In *The Thirteenth International Conference on Learning Representations*, 2025.

Geoffrey Irving, Paul Christiano, and Dario Amodei. AI safety via debate. *arXiv preprint arXiv:1805.00899*, 2018.

Zachary Kenton, Noah Yamamoto Siegel, Janos Kramar, Jonah Brown-Cohen, Samuel Albanie, Jannis Bulian, Rishabh Agarwal, David Lindner, Yunhao Tang, Noah Goodman, and Rohin Shah. On scalable oversight with weak LLMs judging strong LLMs. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=O1fp9nVraj.

Akbir Khan, John Hughes, Dan Valentine, Laura Ruis, Kshitij Sachan, Ansh Radhakrishnan, Edward Grefenstette, Samuel R Bowman, Tim Rocktäschel, and Ethan Perez. Debating with more persuasive LLMs leads to more truthful answers. In *Forty-first International Conference on Machine Learning*, 2024.

Jan Hendrik Kirchner, Yining Chen, Harri Edwards, Jan Leike, Nat McAleese, and Yuri Burda. Prover-verifier games improve legibility of LLM outputs. *arXiv preprint arXiv:2407.13692*, 2024.

Raghav Kulkarni and Avishay Tal. On fractional block sensitivity. *Chicago Journal of Theoretical Computer Science*, 2016(8):1–16, 2016. doi: 10.4086/cjtcs.2016.008.

Jan Leike, David Krueger, Tom Everitt, Miljan Martic, Vishal Maini, and Shane Legg. Scalable agent alignment via reward modeling: a research direction. *arXiv preprint arXiv:1811.07871*, 2018.

Monte MacDiarmid, Benjamin Wright, Jonathan Uesato, Joe Benton, Jon Kutasov, Sara Price, Naia Bouscal, Sam Bowman, Trenton Bricken, Alex Cloud, Carson Denison, Johannes Gasteiger, Ryan Greenblatt, Jan Leike, Jack Lindsey, Vlad Mikulik, Ethan Perez, Alex Rodrigues, Drake Thomas, Albert Webson, Daniel Ziegler, and Evan Hubinger. Natural emergent misalignment from reward hacking in production rl, 2025. URL https://arxiv.org/abs/2511.18397.

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

Tejal Patwardhan, Rachel Dias, Elizabeth Proehl, Grace Kim, Michele Wang, Olivia Watkins, Sim'on Posada Fishman, Marwan Aljubeh, Phoebe Thacker, Laurance Fauconnet, Natalie S. Kim, Patrick Chao, Samuel Miserendino, Gildas Chabot, David Li, Michael Sharman, Alexandra Barr, Amelia Glaese, and Jerry Tworek. Gdpval: Evaluating ai model performance on real-world economically valuable tasks. *ArXiv*, abs/2510.04374, 2025.

Zhihong Shao, Yuxiang Luo, Chengda Lu, Z. Z. Ren, Jiewen Hu, Tian Ye, Zhibin Gou, Shirong Ma, and Xiaokang Zhang. Deepseekmath-v2: Towards self-verifiable mathematical reasoning, 2025. URL https://arxiv.org/abs/2511.22570.

Stephan Wäldchen, Kartikey Sharma, Max Zimmer, and Sebastian Pokutta. Merlin-Arthur classifiers: Formal interpretability with interactive black boxes. *arXiv preprint arXiv:2206.00759*, 2022.

NeurIPS Paper Checklist

1. **Claims**

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

   Answer: [Yes].

   Justification: The abstract and introduction state the protocol contributions, the improvement from average-case to worst-case guarantees, the stronger equilibrium claim, and the black-box lower-bound scope. The formal assumptions and results are developed in the preliminaries, protocol, and lower-bound sections.

   Guidelines:

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

2. **Limitations**

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

   Answer: [Yes].

   Justification: The discussion section states that the lower bounds apply in the black-box human-judgment model and discusses the obstacle of exploiting debaters' computational abilities beyond fractional block sensitivity. The introduction and lower-bound sections also identify the structural assumptions behind the protocol.

   Guidelines:

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

3. **Theory assumptions and proofs**

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

   Answer: [Yes].

   Justification: The paper defines the debate framework, recursive decompositions, and fractional block sensitivity before stating the completeness, soundness, and lower-bound results. Proofs are provided for the stated theorems and supporting lemmas in the main text and Appendices A and B.

   Guidelines:

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

4. **Experimental result reproducibility**

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

   Answer: [N/A].

   Justification: The paper is theoretical and does not report experimental results.

   Guidelines:

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

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

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

   Answer: [N/A].

   Justification: The paper does not include experiments requiring data or code release.

   Guidelines:

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

6. **Experimental setting/details**

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

   Answer: [N/A].

   Justification: The paper does not include experiments, training runs, test splits, or hyperparameters.

   Guidelines:

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

7. **Experiment statistical significance**

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

   Answer: [N/A].

   Justification: The paper does not include empirical experiments or statistical estimates.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.

- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

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

   Answer: [N/A].

   Justification: The paper does not report computational experiments.

   Guidelines:

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

9. **Code of ethics**

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

   Answer: [Yes].

   Justification: The authors have reviewed and followed the NeurIPS Code of Ethics.

   Guidelines:

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

10. **Broader impacts**

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

    Answer: [Yes].

Justification: The broader impacts paragraph in Section 6 discusses the potential benefit to scalable oversight and the risk of unwarranted confidence in incorrect or harmful outputs when the formal assumptions fail.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

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

Answer: [N/A].

Justification: The paper does not release data, models, or other artifacts with high misuse risk.

Guidelines:

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

12. **Licenses for existing assets**

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

Answer: [N/A].

Justification: The paper does not use existing datasets, codebases, models, or other external assets as research artifacts.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.

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

13. **New assets**

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

    Answer: [N/A].

    Justification: The paper does not introduce or release new datasets, code packages, models, or similar assets.

    Guidelines:

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

14. **Crowdsourcing and research with human subjects**

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

    Answer: [N/A].

    Justification: The paper does not involve crowdsourcing or human-subjects research.

    Guidelines:

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

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

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

    Answer: [N/A].

    Justification: The paper does not involve crowdsourcing or human-subjects research.

Guidelines:

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

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A].

    Justification: The core method and results are theoretical and do not use LLMs as an important, original, or non-standard methodological component.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.

A Proof of Theorem 4.2

Theorem 4.2. *Suppose $A(x) \neq L(x)$. Then, regardless of Alice's strategy,*

$$\mathbb{P}[\text{Alice loses}] \;\geq\; \rho^{-d}.$$

Proof. Let $\hat{x}^k$ be the problem at level $k \in \{d, d-1, \ldots, 0\}$ during the execution of the FBS debate with $\hat{x}^d := x$. At level k, let

$$a_k := A(\hat{x}^k), \qquad g_k := L(\hat{x}^k),$$

and let E_k be the event that $a_k \neq g_k$. For $k \in \{d, \ldots, 0\}$, define the event

$$\begin{aligned} W_k := \{&\text{Alice loses before the recursion reaches level } k\} \\ &\cup \{\text{the recursion reaches level } k \text{ and } E_k \text{ holds}\}. \end{aligned}$$

Since $A(x) \neq L(x)$, we have $\mathbb{P}[W_d] = 1$. We claim that for each $k \geq 1$, conditioned on any transcript through level k for which W_k holds,

$$\mathbb{P}\big[W_{k-1} \mid \text{transcript through level } k\big] \;\geq\; \frac{1}{\rho}. \tag{1}$$

If Alice has already lost, or if she fails the consistency or the stability check at level k, then W_{k-1} holds trivially. So assume the recursion reaches level k with E_k and both checks pass. Let y and z be the sub-problems and answers proposed by Alice at level k, and let $\boldsymbol{g} \in \{0,1\}^q$, $\boldsymbol{g}_i := L(y_i)$, be the correct answers to y. Define $B := \{i \in [q] : z_i \neq \boldsymbol{g}_i\}$. Then $B \neq \varnothing$: otherwise $z = \boldsymbol{g}$, and the consistency check would give $a_k = M_D(\hat{x}^k, y, z) = g_k$, contradicting E_k. Moreover, B is sensitive for $M_D(\hat{x}^k, y, \cdot)$ at z, because $z^B = \boldsymbol{g}$ and hence

$$M_D(\hat{x}^k, y, z^B) = M_D(\hat{x}^k, y, \boldsymbol{g}) = g_k \neq a_k = M_D(\hat{x}^k, y, z).$$

Let t be the optimal dual solution computed by Bob in the stability check. By dual feasibility, $\sum_{i \in B'} t_i \geq 1$ for every sensitive block B', and since the check passed, $\sum_i t_i = \mathrm{fbs}_z\big(M_D(\hat{x}^k, y, \cdot)\big) \leq \rho$. Thus the index I chosen by Bob satisfies

$$\mathbb{P}[I \in B] = \sum_{i \in B} \frac{t_i}{\sum_j t_j} \;\geq\; \frac{1}{\sum_j t_j} \;\geq\; \frac{1}{\rho}.$$

If $I \in B$, then $a_{k-1} = z_I \neq \boldsymbol{g}_I = g_{k-1}$, i.e., E_{k-1} occurs and W_{k-1} holds. This proves (1).

As W_k is determined by the transcript through level k, averaging (1) over all transcripts realizing W_k gives $\mathbb{P}[W_{k-1}] \geq \mathbb{P}[W_{k-1} \cap W_k] \geq \rho^{-1}\,\mathbb{P}[W_k]$, and hence

$$\mathbb{P}[W_0] \;\geq\; \rho^{-d}\;\mathbb{P}[W_d] \;=\; \rho^{-d}.$$

Finally, W_0 implies that Alice loses: either she has already lost, or the recursion reaches level 0 and her claimed answer disagrees with $\mathcal{O}$. Hence $\mathbb{P}[\text{Alice loses}] \geq \rho^{-d}$. □

B Proof of Theorem 5.1

We retain the setup of Section 5, including the assumption that the leaf queries are pairwise distinct. The proof proceeds in two steps. First, we bound the checker's distinguishing advantage in terms of the fractional block sensitivity of the Boolean function computed by the entire tree. We then show that this global sensitivity is at least ρ^d when every local sensitivity is at least ρ. We begin by defining the global function and its sensitivity.

Notation. For a leaf ℓ, let q_ℓ be its base-case oracle query. Let

$$Q(\mathcal{T}) := \{q_\ell : \ell \text{ is a leaf of } \mathcal{T}\}$$

be the set of leaf-query strings. The decomposition tree $\mathcal{T}$ induces a Boolean function

$$F_{\mathcal{T}} : \{0,1\}^{Q(\mathcal{T})} \to \{0,1\}$$

by assigning each leaf ℓ the variable indexed by q_ℓ and evaluating the tree bottom-up.

The reference oracle $\mathcal{O}_Y$ induces an assignment $\alpha \in \{0,1\}^{Q(\mathcal{T})}$, defined by $\alpha_{q_\ell} \coloneqq \mathcal{O}_Y(q_\ell)$ for every leaf ℓ. Let

$$\rho_{\mathcal{T}} \coloneqq \mathrm{fbs}_\alpha(F_{\mathcal{T}}),$$

that is, the fractional block sensitivity of the whole decomposition tree $\mathcal{T}$ at the reference assignment α.

We first prove an auxiliary lemma on fractional block sensitivity.

Lemma B.1 (A light sensitive block)**.** *Let $f : \{0,1\}^I \to \{0,1\}$ be a Boolean function on a finite coordinate set I, let $x \in \{0,1\}^I$, and let $R = \mathrm{fbs}_x(f) > 0$. For every distribution p on I, there is a sensitive block $H \in \mathcal{S}_x(f)$ such that*

$$\sum_{i \in H} p_i \le \frac{1}{R}.$$

Proof. Let $\{w_H\}_{H \in \mathcal{S}_x(f)}$ be an optimal primal solution for $\mathrm{fbs}_x(f)$. Thus $\sum_H w_H = R$ and $\sum_{H \ni i} w_H \le 1$ for every $i \in I$. Sample a sensitive block $\mathbf{H}$ with probability proportional to w_H. Then

$$\mathbb{E}\left[\sum_{i \in \mathbf{H}} p_i\right] = \sum_{i \in I} p_i \frac{\sum_{H \ni i} w_H}{\sum_H w_H} \le \sum_{i \in I} p_i \frac{1}{R} = \frac{1}{R}.$$

Therefore, some block H in the support satisfies $\sum_{i \in H} p_i \le 1/R$. □

Recall that Acc_Y and Acc_N denote the events that the adaptive checker $\mathcal{V}$ accepts under $\mathcal{O}_Y$ and $\mathcal{O}_N$, respectively. We show that $\rho_{\mathcal{T}}$ controls the maximal gap between $\mathbb{P}[\mathsf{Acc}_Y]$ and $\mathbb{P}[\mathsf{Acc}_N]$.

Theorem B.2 (Collapsed-function lower bound)**.** *Assume $\rho_{\mathcal{T}} > 0$. For every m-query adaptive checker $\mathcal{V}$ for $\mathcal{T}$, there is an oracle $\mathcal{O}_N$ such that*

$$\mathrm{val}_{\mathcal{O}_N}(\mathrm{root}) \neq \mathrm{val}_{\mathcal{O}_Y}(\mathrm{root})$$

and

$$|\mathbb{P}[\mathsf{Acc}_Y] - \mathbb{P}[\mathsf{Acc}_N]| \le \min\left\{1, \frac{m}{\rho_{\mathcal{T}}}\right\}.$$

Proof. For any leaf query $s \in Q(\mathcal{T})$, let

$$a_s = \mathop{\mathbb{P}}_{\mathcal{O}_Y}[\mathcal{V} \text{ queries } s].$$

Then

$$\Gamma \coloneqq \sum_{s \in Q(\mathcal{T})} a_s = \mathbb{E}_{\mathcal{O}_Y}\left[\#\{s \in Q(\mathcal{T}) : \mathcal{V} \text{ queries } s\}\right] \le m,$$

since the number of distinct leaf-query strings queried is at most the total number of oracle queries.

If $\Gamma > 0$, set $p_s = a_s/\Gamma$; if $\Gamma = 0$, take any distribution p on $Q(\mathcal{T})$. By Lemma B.1, there is a sensitive block H for $F_{\mathcal{T}}$ at α with $p(H) \coloneqq \sum_{s \in H} p_s \le 1/\rho_{\mathcal{T}}$. Define $\mathcal{O}_N$ by flipping $\mathcal{O}_Y$ on H and leaving all other oracle strings unchanged. Since H is sensitive, the root value flips. Moreover,

$$\sum_{s \in H} a_s \le \Gamma \le \frac{m}{\rho_{\mathcal{T}}},$$

with the case $\Gamma = 0$ being trivial.

Couple the two executions with the same random tapes for $\mathcal{V}$ and Alice. Since the oracles agree outside H, the two executions have identical transcripts until $\mathcal{V}$ first queries a string in H. Therefore the outputs can differ only if $\mathcal{V}$ queries some string in H under $\mathcal{O}_Y$. By the union bound,

$$|\mathbb{P}[\mathsf{Acc}_Y] - \mathbb{P}[\mathsf{Acc}_N]| \le \mathop{\mathbb{P}}_{\mathcal{O}_Y}[\mathcal{V} \text{ queries } H] \le \sum_{s \in H} a_s \le \frac{m}{\rho_{\mathcal{T}}}.$$

The left-hand side is also at most 1. □

We next lower bound $\rho_{\mathcal{T}}$ in terms of the local fractional block sensitivities at the internal nodes.

Lemma B.3. *Let $\rho \geq 1$, and suppose every internal node u satisfies*

$$\mathrm{fbs}_{\boldsymbol{z}^u}(\phi_u) \geq \rho,$$

where $\boldsymbol{z}^u \in \{0,1\}^q$ is the vector of child values induced by $\mathcal{O}_Y$. Then $\rho_{\mathcal{T}} \geq \rho^d$.

Proof. By LP duality, it is enough to show that every distribution p on $Q(\mathcal{T})$ admits a sensitive block for $F_{\mathcal{T}}$ at α of p-mass at most ρ^{-d}. We prove this by induction on d.

For $d = 0$ the claim is immediate. For $d \geq 1$, let r be the root and let $\mathcal{T}_i$ be the subtree rooted at its ith child. Put $p_i = p(Q(\mathcal{T}_i))$. Because the leaf queries are pairwise distinct, the sets $Q(\mathcal{T}_i)$ are disjoint, so $(p_1, \ldots, p_q)$ is a distribution on $[q]$. By Lemma B.1 applied to ϕ_r at $\boldsymbol{z}^r$, there is a sensitive block $J \subseteq [q]$ with $\sum_{i \in J} p_i \leq 1/\rho$.

For each $i \in J$, apply the induction hypothesis inside $\mathcal{T}_i$ to the conditional distribution $p^{(i)}$ on $Q(\mathcal{T}_i)$, choosing an arbitrary distribution if $p_i = 0$. This gives a sensitive block $H_i \subseteq Q(\mathcal{T}_i)$ with $p^{(i)}(H_i) \leq \rho^{-(d-1)}$. Let $H = \bigcup_{i \in J} H_i$. Flipping H flips exactly the child values in J, and J is sensitive for ϕ_r; hence H is sensitive for $F_{\mathcal{T}}$. Finally,

$$p(H) = \sum_{i \in J : p_i > 0} p_i \, p^{(i)}(H_i) \leq \rho^{-(d-1)} \sum_{i \in J} p_i \leq \rho^{-d}.$$

□

We now return to the main lower bound.

Theorem 5.1. *Let $\rho \geq 1$, and suppose every internal node u satisfies*

$$\mathrm{fbs}_{\boldsymbol{z}^u}(\phi_u) \geq \rho,$$

where $\boldsymbol{z}^u \in \{0,1\}^q$ is the vector of child values induced by $\mathcal{O}_Y$. Then for every m-query adaptive checker $\mathcal{V}$ for $\mathcal{T}$, there is an oracle $\mathcal{O}_N$ such that

$$\mathrm{val}_{\mathcal{O}_N}(\mathrm{root}) \neq \mathrm{val}_{\mathcal{O}_Y}(\mathrm{root})$$

and

$$|\mathbb{P}[\mathsf{Acc}_Y] - \mathbb{P}[\mathsf{Acc}_N]| \leq \min\{1, m\rho^{-d}\}.$$

Proof. By Lemma B.3, $\rho_{\mathcal{T}} \geq \rho^d \geq 1$. The claim then follows from Theorem B.2, since

$$\min\left\{1, \frac{m}{\rho_{\mathcal{T}}}\right\} \leq \min\{1, m\rho^{-d}\}.$$

□

Consequently, if a checker outputs the correct root value with probability at least $1 - \varepsilon$ under each oracle, for some $\varepsilon < 1/2$, then $1 - 2\varepsilon \leq m\rho^{-d}$ and hence $m \geq (1 - 2\varepsilon)\rho^d$.

Remark B.4. The ρ^d scale in Theorem 5.1 is the standard composition barrier for fractional certificates. Once $\mathcal{T}$ is fixed, its leaves define the composed Boolean function $F_{\mathcal{T}}$. The composition theory of Gilmer, Saks, and Srinivasan [Gilmer et al., 2016], together with the equivalence between fractional and randomized certificates, recovers the same $\Omega(\rho^d)$ randomized-query lower bound in the independent-leaf case. Theorem 5.1 is the corresponding debate-verification form: it keeps the revealed tree fixed, allows arbitrary adaptive oracle queries, and gives the explicit advantage bound $m\rho^{-d}$.