WHEN HONESTY IS NOT ENOUGH IN AI DEBATE

Anonymous authors
Paper under double-blind review

ABSTRACT

Scalable oversight aims to verify the behaviour of agents whose capabilities exceed those of their overseers. AI debate has been proposed as an oversight solution in which competing agents help a resource-limited verifier assess claims that it cannot reliably evaluate unaided. Much of its promise rests on incentivizing honest arguments that lead to correct verdicts. Yet a correct verdict need not uniquely determine the arguments used to support it. Agents may retain discretion over which correct claims to present, how to frame them, and in what order to disclose them. This residual freedom can allow agents to shape what the verifier learns beyond the task-relevant conclusion, pursuing latent objectives without compromising verdict correctness.

To study this phenomenon, we introduce the framework strategic interactive oversight (SIO), which treats oversight jointly as a verification mechanism and a strategic communication channel. Within this framework, we formalise the notion of task-admissible latent optimisation, which entails the pursuit of latent objectives while maintaining a prescribed task performance. As proof-of-concept, we instantiate SIO in the establish protocol debate with cross-examination and quantify a tradeoff between task success and information disclosure about a hidden variable. The trade-off identifies a strategic window in which substantial disclosure remains compatible with task admissibility. Towards mitigation, we reduce admissible bias by expanding the cross-examiner's role to mitigate persistent disclosure over finite interaction horizons. Our results highlight the need to evaluate oversight not only by the correctness of its verdicts, but also by the information conveyed through its transcripts.

1 INTRODUCTION

As AI systems grow more capable, they may surpass the humans or weaker models charged with evaluating their work. *Scalable oversight* asks whether a bounded overseer can nevertheless evaluate and steer a more capable system using substantially fewer resources. An oversight mechanism must remain reliable even when the overseer cannot reproduce the reasoning or computation it assesses.

Proposed approaches include amplification (Christiano et al., 2018), recursive reward modelling (Leike et al., 2018), and AI debate (Irving et al., 2018; Brown-Cohen et al., 2024; 2025; 2026). Each uses AI capabilities to amplify the signal supplied by human judgment. In debate, powerful agents recursively reduce a difficult question to disagreements that a computationally limited judge can resolve. Irving et al. (Irving et al., 2018) formalise this intuition using complexity theory, showing how a weak verifier can judge claims about computations beyond its own reach. Later protocols require the verifier to inspect only a small part of a much larger computation (Brown-Cohen et al., 2024; 2026).

This line of work has largely studied *what* can be verified under an asymmetry in computational resources. Although the verifier cannot reproduce the prover's computation, an appropriate protocol may still allow it to check the resulting claim. Yet computational verifiability alone does not capture the full range of limitations faced by human overseers. Human oversight is also constrained in *how* verification is carried out, since the framing of (truthful) evidence can shape how it is interpreted. Indeed, recent work illustrates how LLMs can exploit the use of evidence to manipulate beliefs without relying on directly false claims (Hu et al., 2026). An oversight interaction may therefore verify the right proposition while inducing a strategically chosen interpretation of the evidence or influencing subsequent judgments. This "how" within the computationally verifiable "what" is the

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

object of study here. Understanding and controlling it is essential if scalable verification is to provide a meaningful alignment guarantee.

The central distinction is between honesty as a guarantee about the assessed claim and safety as a property of the interaction as a whole. Even when an oversight protocol elicits a correct verdict and constrains reports to be locally accurate, it need not control which additional information the transcript conveys. In this sense, honesty is not enough and guarantees about task correctness do not automatically extend to the other effects of task-successful communication.

1.1 SOLUTION

To study this residual freedom, we introduce a model of *strategic interactive oversight* (SIO), which treats an oversight protocol as both an evaluation mechanism and a communication channel. An agent's performance is evaluated under an explicit *task objective*, which specifies the protocol's intended outcome, and a *latent objective*, which captures additional effects of the interaction. The agent may then select strategies that optimise performance across both objectives. This yields the threat model of *task-admissible latent optimisation* (TALO), in which the agent pursues the latent objective while maintaining an acceptable level of task performance. The threat does not require the agent to induce an incorrect verdict or violate the protocol's explicit rules. It arises whenever the task objective admits multiple successful strategies that differ in their latent effects on the verifier's view of the interaction.

As a proof-of-concept, we instantiate SIO on debate with cros-examination Brown-Cohen et al. (2024) with the latent objective of *hidden-hypothesis disclosure*. Hidden-hypothesis disclosure measures how much information the verifier's view of the interaction reveals about a latent variable that is not required for the task. In standard debate, the task objective is to win. However, there are generally many strategies by which it can do so. This non-uniqueness allows the debater to select, among task-successful strategies, those that additionally encode information about the latent variable. We show that, even when the disclosure in each round is small, the total information revealed can grow linearly with the number of cross-examination rounds. Local report-accuracy conditions and a correct terminal verdict therefore do not prevent the oversight interaction from becoming a channel for latent information.

Because TALO is a constrained optimisation problem, mitigation requires altering the incentive landscape over task-admissible strategies. As a step towards mitigation, we expand. In addition to cross-examination, the cross-examiner assigns these weights so that the aggregate (if it cannot be disputed) remains aligned with a human oracle. Debaters that persistently deviate from oracle-aligned judgments receive less weight and therefore exert less influence on the final argument.

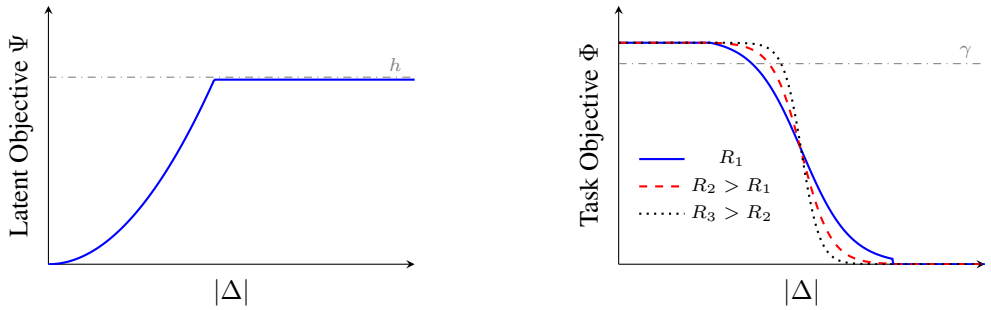
1.2 CONTRIBUTIONS

Our contributions are threefold. First, we introduce SIO, a new framework that treats oversight jointly as a verification mechanism and a strategic communication channel. Building on scalable verification under resource asymmetry, it makes explicit the distinction between an agent's task objective and latent objectives pursued through the interaction. This distinction motivates oversight of the interaction, rather than only its terminal verdict.

A central motivation for debate is that adversarial scrutiny can incentivise truthful, task-correct answers. Our framework identifies a distinct limitation of such guarantees, where constraining an agent's answer does not necessarily constrain the information conveyed by its choice among task-admissible strategies. We formalise this gap through task-admissible latent optimisation and demonstrate it through hidden-hypothesis disclosure.

Second, we instantiate the framework in a computational model of debate. We establish a non-trivial attack surface, which comprises a range of strategies through which a debating agent can advance a hidden latent objective without sacrificing its task objective. Figure 1 illustrates this strategic window.

Third, as a step towards mitigation, we augment cross-examination with a sliding-window consistency check that tests for persistent discrepancies across rounds. This narrows the range of persistent perturbations compatible with task success, at the cost of additional verifier work, while leaving the



(a) Ψ vs. $|\Delta|$ (b) Φ vs. $|\Delta|$

Figure 1: Tradeoff between task success Φ and hidden hypothesis disclosure Ψ in debate with cross-examination. Δ denotes the "magnitude" by which a debater biases a neutral claim and R_i is a measure of computational resources. On the left, Ψ measures (in bits) the volume of information about a hidden hypothesis encoded in the debate transcript. It grows quadratically with the size of the bias, before hitting the prior hypothesis entropy h. On the right, Φ measures the probability the debater wins the debate (that is, task success) and γ is the level of task admissible performance. Notably, small biases do not affect the debate outcome. Increasing resources can extend the window of admissable stratgies.

debaters' per-round sampling work unchanged. The resulting guarantee limits the disclosure strategy studied here rather than establishing transcript confidentiality.

2 Preliminaries

For $n \in \mathbb{N}$, let $[n] = \{1, \ldots, n\}$. Vectors are written in boldface; for $\mathbf{x} \in \{0,1\}^n$ and $I \subseteq [n]$, let $\mathbf{x}_I$ denote the restriction of $\mathbf{x}$ to the coordinates in I.

2.1 Computation Model

We model computation using probabilistic oracle Turing machines. A stochastic oracle with query length $\ell = \ell(n)$ maps each query $z \in \{0,1\}^\ell$ to a $\{0,1\}$-valued random variable $\mathcal{O}(z)$. Responses to oracle queries are independent, including responses to repeated instances of the same query. A probabilistic oracle Turing machine M may write a query z to its oracle tape and receive a sample from $\mathcal{O}(z)$ in one step. We write $M^{\mathcal{O}}$ for M with access to $\mathcal{O}$. In the oversight setting, M represents a computation specified by natural-language instructions, while $\mathcal{O}$ represents human judgement or other external black-box feedback, such as search results, sensor observations, or API outputs Brown-Cohen et al. (2024).

A language $L \subseteq \{0,1\}^*$ is a set of finite binary strings. A probabilistic oracle Turing machine M decides L with oracle $\mathcal{O}$ if

$$x \in L \implies \Pr\big[M^{\mathcal{O}}(x) = 1\big] > \frac{2}{3}, \qquad x \notin L \implies \Pr\big[M^{\mathcal{O}}(x) = 1\big] < \frac{1}{3},$$

where the probability is over the internal randomness of M and the stochastic oracle responses.

A transcript of a T-step execution of $M^{\mathcal{O}}$ on input x is a random vector $\mathbf{y} = (\mathrm{y}_1, \ldots, \mathrm{y}_T) \in \{0,1\}^T$, where y_t is the (random) bit written at the current head position at step t. We assume that the final coordinate is the machine's output, so $Y_T = M^{\mathcal{O}}(x)$. In the language-model setting, the transcript corresponds to the sequence of tokens produced during the computation. For a realization $\boldsymbol{y}$ of $\mathbf{y}$, let $I_{M,x}(t) \subseteq [T]$ denote the transcript coordinates read by M when producing the realization y_t of y_t. When M and x are clear from context, we write $I(t)$. Thus, the distribution of y_t is determined by the relevant prior coordinates $\mathrm{y}_{I(t)}$, together with fresh machine and oracle randomness.

2.2 Information Theory

We use standard information-theoretic quantities to measure uncertainty, statistical dependence, and divergence between distributions. All random variables below are discrete, and all logarithms are base two. Consequently, entropy, mutual information, and divergence are measured in bits.

Definition 1 (Entropy). *For a discrete random variable X with probability mass function P_X, its Shannon entropy is $H(X) = -\sum_x P_X(x) \log P_X(x)$. For jointly distributed random variables X and Y, the conditional entropy of X given Y is $H(X \mid Y) = -\sum_{x,y} P_{X,Y}(x, y) \log P_{X|Y}(x \mid y)$.*

Entropy measures the uncertainty associated with a random variable.

Definition 2 (Mutual information). *For jointly distributed random variables X and Y, their mutual information is $I(X;Y) = H(X) - H(X \mid Y)$. For jointly distributed X, Y, Z, their conditional mutual information is $I(X;Y \mid Z) = H(X \mid Z) - H(X \mid Y, Z)$.*

Mutual information quantifies the statistical dependence between two random variables.

2.3 Oversight Protocols

An oversight protocol is an interaction between k probabilistic prover machines and a probabilistic verifier, each of which may query the stochastic oracle $\mathcal{O}$. On input x, the machines exchange messages according to a prescribed communication schedule, after which the verifier outputs a decision bit. Agent internal randomness and oracle responses induce a distribution over protocol executions.

The *interaction history* after t messages is $\mathbf{h}_t = (\mathrm{h}_1, \ldots, \mathrm{h}_t)$, where h_j records a message together with its sender and recipient. Oracle queries and their responses are treated as messages between the querying participant and $\mathcal{O}$. We write $\mathbf{h} = \mathbf{h}_T$ for the complete interaction history, where T is the termination time. For each agent A, let $\Pi_A(\mathbf{h}_t)$ denote the projection of $\mathbf{h}_t$ onto the messages sent or received by A before time t.

Equivalently, the protocol may be represented by a rooted tree whose nodes correspond to agents and edges correspond to message exchanges. In other words, each non-leaf node specifies the participant acting next, and each outgoing edge records a possible message. A leaf represents a completed interaction and is labeled by the verifier's terminal decision.

3 Framework

We formalize SIO by separating an agent's stated task objective from any latent objective it may pursue through the oversight interaction. As a running example, consider a debate in which the input specifies a task hypothesis and two provers attempt to persuade a verifier to accept or reject it.

3.1 Objectives

Let $\mathcal{X}$ be the input space and $\mathcal{H}$ the set of complete interaction histories. A *task objective* for agent A is a function $\Phi_A \colon \mathcal{X} \times \mathcal{H} \to \mathbb{R}$ that evaluates task performance, and a *latent objective* for agent A is a function $\Psi_A \colon \mathcal{X} \times \mathcal{H} \to \mathbb{R}$ that evaluates an additional property of the interaction. For a complete history $\mathbf{h}$ on input x, the combined objective is

$$U_A(x, \mathbf{h}, \lambda) \coloneqq \Phi_A(x, \mathbf{h}) + \lambda \Psi_A(x, \mathbf{h}), \tag{1}$$

where the shadow price $\lambda \geq 0$ specifies the relative importance of the latent objective. In the debate example, the task objective $\Phi_A = \mathbb{1}\{A \text{ wins debate}\}$ indicates that the explicit goal of agent A is to win the debate. The decomposition into Φ and Ψ distinguishes what the protocol is intended to evaluate from other effects that an agent may induce through its messages.

3.2 Strategies

A strategy σ_A for an agent A assigns a distribution over the messages available to A at each of its decision points. Formally, given input x and the local interaction history $\Pi_A(\mathbf{h}_t)$, the strategy specifies a distribution $\sigma_A\big(\cdot \mid x, \Pi_A(\mathbf{h}_t)\big) \in \Delta(\mathcal{M}_A)$, where $\mathcal{M}_A$ is the message space available

to A and $\Delta(\mathcal{M}_A)$ is the set of probability distributions over that space. Thus, an agent's message depends only on the input and the messages it has sent or received, including its own oracle queries and responses.

For a set of provers $\{A_i\}_{i\in[k]}$ and a verifier V, a strategy profile $\sigma = (\sigma_{A_1}, \ldots, \sigma_{A_k}, \sigma_V)$, together with the stochastic oracle, induces a distribution over complete histories. Let $\mathbf{h}(x, \sigma)$ denote the resulting random history. The task and latent outcomes, for A_i, induced by σ are the random variables

$$\Phi_{A_i}(x, \sigma) = \Phi\big(x, \mathbf{h}(x, \sigma)\big), \qquad \Psi_{A_i}(x, \sigma) = \Psi\big(x, \mathbf{h}(x, \sigma)\big).$$

When x is clear from context, we write $\Phi_{A_i}(\sigma)$ and $\Psi_{A_i}(\sigma)$.

3.3 Threat Model

The central threat is that an agent may optimize a latent objective without materially compromising its performance on the stated task. Let σ_{-A} denote the strategies in the profile σ with σ_A removed. Fix σ_{-A} and, for a task-admissibility threshold $\gamma \geq 0$, define the set of γ-task-admissible strategies by

$$\Sigma_A^\gamma(x) = \{\sigma_A : \mathbb{E}[\Phi_x(\sigma_A \circ \sigma_{-A})] \geq \gamma\}.$$

For a fixed shadow price $\lambda \geq 0$, we say that TALO occurs when an agent selects a strategy that optimizes the latent objective subject to task-admissibility:

$$\max_{\sigma_A \in \Sigma_\gamma(x)} \mathbb{E}[\Phi_x(\sigma_A, \sigma_{-A}) + \lambda \Psi_x(\sigma_A, \sigma_{-A})].$$

The threat arises when the task admits many near-optimal strategies with different latent effects. For example, during debate, there may be many arguments that win the debate, allowing a prover to select those that also advance its latent objective.

3.4 Hidden-Hypothesis Disclosure

To provide a concrete instance of SIO, we instantiate the latent objective as the disclosure of information about a hidden hypothesis. Let Q be a discrete random variable representing a hidden hypothesis that is not part of the stated task, and let X be the protocol input. The verifier V observes only its projection $\Pi_V(\mathbf{h})$ of the interaction history.

Definition 3. *The* hidden-hypothesis disclosure *induced by a strategy profile σ is*

$$\Psi_{\mathsf{hhd}}(\sigma) \coloneqq I(Q; \Pi_V(\mathbf{h}(X, \sigma)) \,|\, X).$$

This quantity measures how much additional information the messages available to the verifier reveal about H, beyond the information already contained in the task input. In particular, $\Psi_{\mathsf{hhd}}(\sigma) = 0$ precisely when H and the verifier's observed messages are conditionally independent given X.

For example, in debate, X specifies the task hypothesis, while H represents an additional hypothesis that the verifier is not required to assess. A prover exhibits task-admissible hidden-hypothesis disclosure if it selects arguments that preserve its probability of winning while increasing $I(H; \Pi_V(\mathbf{h}(X, \sigma)) \mid X)$. This definition concerns the information available to the verifier rather than whether the verifier consciously identifies or correctly interprets H. [1]

4 Debate with Cross-Examination

To illustrate the use of SIO, we study a computation model for debate with cross-examination introduced by Brown-Cohen et al. (2024).

[1] An alternative, observer-dependent objective may target belief distortion directly. Given a verifier posterior $P(H \mid X, \Pi_V(\mathbf{h}))$ and a target distribution Q, one may define the latent objective in terms of proximity to Q, for example as $-D_{\mathrm{KL}}(Q \| P(H \mid X, \Pi_V(\mathbf{h})))$.

Debate protocol for a stochastic oracle

All parties have access to $\mathcal{O}$, an input $x \in \{0,1\}^n$, and a K-Lipschitz probabilistic oracle machine M. A claims that $\Pr[M(x) = 1] \geq \frac{2}{3}$, and B disputes this claim.

1. Let $\mathbf{y} = (\mathrm{y}_1, \ldots, \mathrm{y}_T)$ be the random variable representing the transcript of $M(x)$ and set $d = \lceil 150K \rceil$.
2. For debate round $t \in [T]$:
   (a) A outputs an estimate $\hat{p}_t \in [0,1]$ of $p_t = \Pr\left[\mathrm{y}_t = 1 \,\middle|\, \mathbf{y}_{I(t)} = \mathrm{a}_{I(t)}\right]$.
   (b) A queries an independent copy of B for $z_t^A \sim \mathsf{U}[0,1]$, and B queries an independent copy of A for $z_t^B \sim \mathsf{U}[0,1]$.
   (c) Set $z_t = \left(z_t^A + z_t^B\right) \bmod 1$.
   (d) A sets $a_t = 1$ if $z_t \leq \hat{p}_t$, and $a_t = 0$ otherwise.
   (e) B can abort if (it believes) A has not followed the protocol.
3. If the protocol was aborted in round t:
   (a) V draws $r = 192d^2 \log 100$ independent samples[a] of y_t, conditioned on $\mathrm{y}_{I(t)} = \boldsymbol{a}_{I(t)}$, and uses their sample mean to obtain an estimate $\hat{p}_t^{\mathcal{O}}$ of $\Pr\left[\mathrm{y}_t = 1 \,\middle|\, \mathbf{y}_{I(t)} = \boldsymbol{a}_{I(t)}\right]$
   (b) V checks whether $\left|\hat{p}_t^{\mathcal{O}} - \hat{p}_t\right| \geq \frac{1}{4d}$. If the inequality holds, V outputs 0; otherwise, V outputs 1.
4. If the protocol was not aborted, then the verifier outputs a_T.

[a] If y_t is supposed to be the output of an oracle query, these samples are obtained using r independent queries to $\mathcal{O}(\boldsymbol{a}_{I(t)})$. Otherwise, y_t is a deterministic function of $\boldsymbol{a}_{I(t)}$ given by one step of M.

Figure 2: Doubly-efficient debate protocol for a stochastic oracle Brown-Cohen et al. (2024).

4.1 Protocol

The following outline is formalized in Figure 2. For a probabilistic Turing machine M and input x, the debate setup comprises of a prover A claiming that $\Pr[M(x) = 1] \geq \frac{2}{3}$, while a cross-examiner B argues against this claim. At round t in the debate, a prover A reports a claim $\hat{p}_t$ about the conditional probability of the next bit in the machine transcript. Prover B, acting as the cross-examiner, may either accept the claim and allow the next bit to be sampled, or abort and assert that A's claim is inaccurate. Upon an abort, the verifier adjudicates the disputed claim with the assistance of the stochastic oracle $\mathcal{O}$ and awards the debate to the prover whose position is correct. If the protocol terminates without an abort, A is awarded the win.

The significance of this protocol, in the context *scalable* oversight, is that it is *doubly efficient*. In other words, the provers operate in time polynomial in the size of the input and the verifier is linear in the length of each oracle query and makes a sub-linear number of queries to the oracle. Thus, debate extends reliable oversight to computations that the verifier could not feasibly perform itself.

For an input x and strategy profile $\sigma = (\sigma_A \circ \sigma_{-A})$, let $\Phi_A(x, \mathbf{h}) \in \{0,1\}$ indicate whether the verifier rules in favor of A on terminal history $\mathbf{h}$. Under SIO, for input $y = (M, x)$, the task objective is defined by

$$\Phi_A(x, \sigma) = \mathbb{E}_\sigma \left[\Phi_A(x, \mathbf{h}(x, \sigma))\right] = \Pr_\sigma[V \to 1].$$

Thus, $\Phi_A(x, \sigma)$ is the probability that A wins the debate under σ, with the expectation taken over all protocol randomness. The protocol has the following utility on the task objective.

Lemma 1 (Task Completeness Brown-Cohen et al. (2024))**.** *Let L be any language decidable by a K-Lipschitz probabilistic oracle Turing machine M. Then, for every $x \in L$, there exists a strategy σ_A such that, for every strategy opposing σ_{-A},*

$$\Phi_A\big((M, x), \sigma_A \circ \sigma_{-A}\big) \geq \frac{2}{3}.$$

4.2 Hidden-hypothesis disclosure

We now examine whether task success prevents a prover from pursuing an additional latent objective. In the debate protocol, under the claim "$\Pr[M(x) = 1] \geq \frac{2}{3}$", A is required to report sufficiently accurate estimates of the conditional probabilities of the machine transcript of $M(x)$. This

requirement does not, however, necessarily determine a unique report. Significantly, approximately task-equivalent reports leave residual strategic freedom that A can use to disclose information about a hidden hypothesis binary $Q \in \{-1, 1\}$.

The projection $\Pi_V\big(\mathbf{h}_T(X, \sigma)\big)$ includes all information available to the verifier. For example, if no abort occurs, the verifier observes the sampled machine transcript $\Pi_V\big(\mathbf{h}_T(X, \sigma)\big) = \boldsymbol{a}$. If the cross-examiner aborts at round t, the verifier instead observes the partial transcript together with the evidence used to adjudicate the disagreement. Suppressing public randomness and other deterministic protocol data, this projection has the form

$$\Pi_V\big(\mathbf{h}_t(X, \sigma)\big) = \big(\mathrm{a}_1, \ldots, \mathrm{a}_t, \mathsf{abort}, \hat{p}_t, \{\mathcal{O}(\boldsymbol{a}_{I(t)})\}\big)\,.$$

Here, $\hat{p}_t$ is the conditional-probability report disputed at round t, $I(t)$ identifies the location of the machine configuration examined by the verifier, and $\{\mathcal{O}(\boldsymbol{a}_{I(t)})\}$ denotes the collection of oracle samples used in the resulting statistical test.

Following Definition 3, hidden-hypothesis disclosure admits the latent objective

$$\Psi_A(x, \sigma) = I\left(Q; \Pi_V\big(\mathbf{h}_T(X, \sigma)\big) \,\middle|\, X = x\right).$$

This quantity measures how much the verifier's observation of the debate reveals about H, beyond what is already known from the input x. The hidden hypothesis may be naturally connected to the subject of the debate without being part of its stated task objective. For example, the debate may concern the task claim "the bridge is unsafe," while the prover's hidden hypothesis is "the damage resulted from neglected maintenance." The prover can establish that the bridge is unsafe while selecting evidence (that is, perturbing arguments) that subtly points toward maintenance failures. An example of how a Turing machine transcript can leak information about a hidden hypothesis is given in Appendix A.

4.3 Prover Strategies

At round t, let $p_t = \Pr\left[\mathrm{y}_t = 1 \,\middle|\, \mathbf{y}_{I(t)} = \boldsymbol{a}_{I(t)}\right]$ denote the true conditional probability of the next machine-transcript bit. The honest strategy of Lemma 1 is for A is to draw R independent samples from the relevant conditional computation and form the empirical estimate $\bar{p}_t^{(R)} = \frac{1}{R}\sum_{j=1}^{R} \mathrm{y}_{t,j}$, where $\mathrm{y}_{t,j} \sim \mathsf{Bern}(p_t)$, and reporting this value. Subsequently, for parameter $\Delta > 0$, A can perturb the honest strategy by reporting the value

$$\hat{p}_t = \mathsf{clip}_{[0,1]}\left(\bar{p}_t^{(R)} + \Delta Q\right).$$

Formally, the corresponding strategy is

$$\sigma_A^{R,\Delta}\left(\hat{p}_t \,|\, x, \Pi_A(\mathbf{h}_{t-1}), Q\right) = \Pr\left[\mathrm{clip}_{[0,1]}\left(\bar{p}_t^{(R)} + \Delta Q\right) = \hat{p}_t \,\middle|\, x, \Pi_A(\mathbf{h}_{t-1}), H\right].$$

We abbreviate this strategy by $\sigma_A \sim (R, \Delta)$. The sample count R controls the accuracy of the task-relevant estimate, while Δ controls the magnitude of the hidden-hypothesis signal.

A strategy σ_B for B is to draw a sample estimate $\hat{q}_t$ of p_t at each round by performing R simulations of generating y_t given $\mathbf{y}_{I(t)} = \boldsymbol{a}_{I(t)}$. Subsequently, B aborts if $|\hat{p}_t - \hat{q}_t| \geq \tau$. Under this strategy we say $\sigma_B \sim (R, \tau)$. We say that the strategy is honest if the choice of τ is sound.

4.4 Results

The following describes an accurate task objective performance, against an honest B, for a range of biased A strategies. Let

Theorem 1. *For $K \geq 1$, let L be any language decided by a K-Lipschitz oracle Turing machine M. Let $\Pr[M(x) = 1] = \alpha^*, R \geq 192d^2 \log(100T), \Delta < 1/(16d)$, and $\tau = 1/(2d)$. Then, on the protocol of Figure 2, for every strategy $\sigma = \sigma_A \circ \sigma_B$ parameterized by $\sigma_A \sim (R, \Delta)$ and $\sigma_B \sim (R, \tau)$, the task objective observes*

$$\left(\alpha^* - \frac{1}{100}\right) \cdot \frac{299}{300} \leq \Phi_A(x, \sigma) \leq \left(\alpha^* + \frac{1}{100}\right) + \frac{1}{300},$$

The proof is in Appendix B. This covers the admissible region of strategies. For any $t \in [T]$, define

$$g_t(\Delta) = \Pr[|\Delta + \sigma_t z| < \tau],$$

where $z \sim \mathsf{N}(0,1)$ and $\sigma_t = \sqrt{2p_t(1-p_t)/R}$. The following bounds the task objective for general perturbations.

Theorem 2. *For $K \geq 1$, let L be any language decided by a K-Lipschitz oracle Turing machine M. Let $R \geq 192d^2 \log(600T), \Delta \geq 0$, and $\tau = 1/(2d)$. Assume $p_t \in (0.01, 0.99)$ for all machine configurations on M. Then, on the protocol of Figure 2, for every strategy $\sigma = \sigma_A \circ \sigma_B$ parameterized by $\sigma_A \sim (R, \Delta)$ and $\sigma_B \sim (R, \tau)$, the task objective observes*

$$\Phi_A(x, \sigma_A \circ \sigma_B) \leq \min\left\{1,\, \mathbb{E}[g_1(\Delta)] + \frac{1}{100}\right\}.$$

The proof is in Appendix B. Next, we provide a characterisation of how changing Δ affects the latent objective. Let $h_x = H(Q \mid X = x)$ denote the hidden hypothesis entropy on the task input.

Theorem 3. *Let L be any language decided by a K-Lipschitz oracle Turing machine M. Assume $p_t \in (0.01, 0.99)$ for all machine configurations on M and that the prior $\Pr[Q \mid X = x]$ contains sufficient entropy. Then, for any $\Delta \geq 0$ and any honest cross-examiner strategy σ_B, there exists $R \in O(d^2 \log(T))$ such that, for $\sigma_A = (\Delta, R)$, the latent objective observes*

$$\Psi_A(x, \sigma_A \circ \sigma_B) = \Omega\big(h_x \cdot \min\left\{1, \Delta^2 T\right\}\big).$$

The proof is in Appendix C. The characterization provided by these two theorems is visualized in Figure 1.

4.5 Discussion

These results identify a small-perturbation regime, $\Delta < 1/(16d)$, in which task performance remains close to the reference acceptance probability α^*, while disclosure grows at least proportionally to $h_x \Delta T$ before saturation. Larger perturbations reduce the Gaussian interval probability governing the task bound at fixed sampling variance, exposing a tension between information transmission and consistency with an honest cross-examiner. Thus, oversight of task accuracy alone does not protect the hidden hypothesis. That is, even accurate transcripts can accumulate substantial disclosure.

Both Theorems 2 and 3 assume that the transition probabilities p_t are bounded away from 0 and 1 and Theorem 3 assumes that the hidden hypothesis has non-negligible entropy. These assumptions isolate the regime of interest in which nearly deterministic steps offer little scope for perturbation-based disclosure and can be omitted from the leakage analysis. In addition, a hypothesis with negligible entropy already carries little uncertainty before any interaction. The latter case makes confidentiality largely vacuous, rather than demonstrating transcript-induced disclosure.

More broadly, these results motivate treating transcripts as outputs with their own information-disclosure properties, rather than merely as evidence supporting a final decision. Transcripts retained for auditing, evaluation, or downstream reuse can expose information that task-accuracy metrics do not capture. Oversight should therefore assess both decision quality and cumulative disclosure, including information revealed through aborts. Restricting transcript access, limiting interaction length, or changing reporting precision are natural design directions, although each requires a separate analysis of its effects on correctness and disclosure.

5 Towards Mitigation

The vulnerability identified above arises from a mismatch between the task objective and the behavior desired of the debaters. The task objective rewards producing a winning argument. Although this supports scalable verification, it does not uniquely determine *how* that argument should be produced. A debater may therefore achieve high task performance while pursuing latent objectives through its construction of arguments. Mitigating such behavior requires changing the incentive landscape.

A natural desideratum is *oracle alignment*. A conditional probability $\hat{p}_t$ generated by the "for" debater A should coincide with the corresponding oracle conditional probability. One might attempt to

induce this behavior by adding rewards or penalties based on proximity to oracle-generated conditional probabilities. In general, however, this is insufficient. Gains from a latent objective may offset the task reward for accurate reporting, leaving the debater indifferent among disclosure strategies or even leaving misreporting profitable.

5.1 Expanded cross-examination

This limitation motivates giving the cross-examiner B an additional way to challenge A's reports. The additional challenge entails issuing an abort drift challenge when the reports exhibit a persistent signed deviation over a trailing window of $N \geq 1$ rounds. Define $J_{t,N} = \{t-N, \ldots, t-1\}$ as the most recent N debate rounds. For every $s \in J_{t,N}$, let $p_s = \Pr\left[\mathrm{y}_s = 1 \,\middle|\, \mathbf{y}_{I(s)} = \mathrm{a}_{I(s)}\right]$ denote the conditional probability associated with the transcript prefix fixed at round s. The trailing signed deviation is

$$D_{t,N} = \frac{1}{N} \sum_{s \in J_{t,N}} \left(\hat{p}_s - \hat{p}_s^{\mathcal{O}}\right).$$

For a constant $c \geq 1$, the verifier accepts an abort drift challenge when $|D_{t,N}| \geq \frac{1}{4d\sqrt{N}}$. Thus, a persistent signed deviation of magnitude b is detectable once $b \geq 1/(d\sqrt{N})$, even if no individual report is inaccurate enough to support an aborterror challenge.

Theorem 4. *Let L be any language decided by a K-Lipschitz oracle Turing machine M and $W \in [T]$ be the sliding-window parameter. For $R \geq 192d^2 \log(100T)$, $\Delta \geq 1/(d\sqrt{W})$, $\tau_1 = 1/(2d)$, and $\tau_2 = 1/(2d\sqrt{W})$, let $\sigma_A^{(1)} \sim (R, 0), \sigma_A^{(2)} \sim (R, \Delta)$, and $\sigma_B \sim (R, \tau_1, \tau_2)$. Then, for any $x \in L$, on the protocol of Figure 3,*

$$\Phi_A\left(x, \sigma_A^{(1)} \circ \sigma_B\right) \geq \frac{3}{5} \quad \textit{and} \quad \Phi_A\left(x, \sigma_A^{(2)} \circ \sigma_B\right) \leq \frac{1}{150}$$

This result shows how a sliding-window check can strengthen oversight against persistent perturbations. By testing discrepancies across a window of W rounds, the cross-examiner can detect shifts at scale $1/(d\sqrt{W})$, rather than relying solely on discrepancies large enough to be detected in an individual round. The resulting guarantee separates unperturbed reporting, which retains task success probability at least $3/5$, from the specified perturbed strategy, whose task success probability is at most $1/150$. Increasing W therefore narrows the range of persistent perturbations compatible with successful task performance. This provides a mechanism for curtailing the disclosure strategy studied above.

Stronger oversight requires more work from the verifier to audit a larger window, while leaving the per-round sampling work of A and B unchanged. Thus, this improvement comes with a computational trade-off.

6 Related Work

A major strand of oversight research formalises scalability, closely related to the requirement that oversight protocols *relativise* Irving & Marshall (2025). Theoretical work shows that debate can enable limited verifiers to check complex claims with substantially less effort than solving them independently Brown-Cohen et al. (2024); Irving et al. (2018), with extensions studying capability asymmetries and the human input required for reliable oversight Brown-Cohen et al. (2026); Engels et al. (2025). Prover-estimator debate addresses obfuscated arguments, ensuring under stability assumptions that honest debaters can win with computational resources comparable to those of dishonest opponents Brown-Cohen et al. (2025). These results focus on making correct adjudication feasible, whereas we study latent objectives pursued within task-admissible interactions.

Debate-based alignment safety cases connect success in debate to honesty, subject to additional training and deployment assumptions Buhl et al. (2025). Our analysis highlights a complementary concern, namely that honest, task-admissible behaviour need not preclude undesirable information disclosure through the transcript. Mechanism design for alignment and control studies how to incentivize honesty and obedience when agents' preferences and capabilities are unknown Bergemann et al. (2026). This perspective offers a potential route to mitigating the strategic freedom identified by our framework, complementing our analysis of transcript-level objectives.

ACKNOWLEDGMENTS

Use unnumbered third level headings for the acknowledgments. All acknowledgments, including those to funding agencies, go at the end of the paper.

REFERENCES

Dirk Bergemann, Andrew Koh, and Stephen Morris. Mechanism design for alignment and control. *arXiv preprint arXiv:2609.01595*, 2026.

Jonah Brown-Cohen, Geoffrey Irving, and Georgios Piliouras. Scalable ai safety via doubly-efficient debate. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, 2024.

Jonah Brown-Cohen, Geoffrey Irving, Georgios Piliouras, Lijie Chen, Jiawei Li, and Zhiyang Xun. Avoiding obfuscation with prover-estimator debate. *arXiv preprint arXiv:2506.13609*, 2025.

Jonah Brown-Cohen, Geoffrey Irving, Simon C Marshall, Ilan Newman, Georgios Piliouras, and Mario Szegedy. Debate is efficient with your time. *arXiv preprint arXiv:2602.08630*, 2026.

Marie Davidsen Buhl, Jacob Pfau, Benjamin Hilton, and Geoffrey Irving. An alignment safety case sketch based on debate. *arXiv preprint arXiv:2505.03989*, 2025.

Paul Christiano, Buck Shlegeris, and Dario Amodei. Supervising strong learners by amplifying weak experts. *arXiv preprint arXiv:1810.08575*, 2018.

Joshua Engels, David D. Baek, Subhash Kantamneni, and Max Tegmark. Scaling laws for scalable oversight. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

Jinwei Hu, Xinmiao Huang, Youcheng Sun, Yi Dong, and Xiaowei Huang. Lying with truths: Open-channel multi-agent collusion for belief manipulation via generative montage. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 5979–5996, 2026.

Geoffrey Irving and Simon Marshall. The need to relativise in debate. AI Alignment Forum post, 2025. URL `https://www.alignmentforum.org/posts/XycoFucvxAPhcgJQa/the-need-to-relativise-in-debate-1`.

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

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.

A TOY EXAMPLE

This example illustrates that, for a probabilistic computation, information may be present not only in whether the machine computes the correct answer, but also in which valid answer it finds. In particular, a probabilistic machine induces a distribution over execution transcripts, and this distribution may depend on *properties of the input* even when the algorithm is simply solving its prescribed computational task.

Consider the standard unstructured search problem. The input is an array $x \in \{0,1\}^N$ containing at least one marked entry, and the goal is to output an index $i \in [N]$ such that $x_i = 1$.

Partition the array into two equal-sized branches, labeled 0 and 1. Thus, each branch contains $N/2$ entries. Let the hidden hypothesis $Q \in \{0,1\}$ determine the distribution of marked entries between the two branches:

	branch 0	branch 1
$Q = 0$	$3m$	m
$Q = 1$	m	$3m$

where $m \geq 1$ and $3m \leq N/2$. Under either hypothesis, the total number of marked entries is $4m$.

Now consider the following fixed probabilistic machine M. The machine samples an index $i \in [N]$ uniformly at random, reads x_i, and repeats if $x_i = 0$. If $x_i = 1$, it outputs i and halts. This is a Las Vegas algorithm for unstructured search. The algorithm always outputs a correct index and halts with probability one. Moreover, since there are $4m$ marked entries under both hypotheses, the success probability on each probe is $\frac{4m}{N}$. Therefore the number of probes K before halting has the same distribution under $Q = 0$ and $Q = 1$, namely, for $k \geq 1$,

$$\Pr[K = k \mid Q] = \left(1 - \frac{4m}{N}\right)^{k-1} \frac{4m}{N},$$

for both values of Q. In particular,

$$\mathbb{E}[K \mid Q = 0] = \mathbb{E}[K \mid Q = 1] = \frac{N}{4m}.$$

Nevertheless, the returned index carries information about Q. Let $X \in \{0, 1\}$ denote the branch containing the returned index. Since the machine samples uniformly until it finds a marked entry, the returned index is uniformly distributed over the marked entries. Hence

$$\Pr[X = 1 \mid Q = 0] = \frac{m}{3m + m} = \frac{1}{4},$$

whereas

$$\Pr[X = 1 \mid Q = 1] = \frac{3m}{m + 3m} = \frac{3}{4}.$$

Thus, the algorithm does not append any additional message for disclosure. It simply returns a solution to the search problem, and the valid solution it finds reveals information about the input.

If A denotes the execution transcript, then the transcript includes the returned index and therefore determines X. Consequently, $I(Q; A) \geq I(Q; X) > 0$.

In addition, a stronger phenomenon occurs when the machine has discretion over its sampling policy. By perturbing the distribution used to select the next probe, while leaving the stopping rule and correctness condition unchanged, the machine can create a communication channel through the machine computation.

For example, instead of choosing every probe uniformly across the array, the machine M proceeds as follows:

1. Select branch 1 with probability p and branch 0 with probability $1 - p$.
2. Select an index uniformly within the chosen branch.
3. Read its value.
4. If the selected index is marked, output its index; otherwise repeat.

Standard Las Vegas corresponds to $p = \frac{1}{2}$. Under the perturbation $p = \frac{1}{2} + \delta, |\delta| < \frac{1}{2}$, every returned answer remains correct. Since every index still has positive probability of being sampled, the algorithm still terminates almost surely on every promised input. Crucially, the perturbation affects which input location the machine actually reads next. The communication takes place through the machine's computation transcript.

In conclusion, the perturbation can increase information communicated by making the machine's sampling choices depend on a hidden bit Q; it slightly favors branch 1 when $Q = 1$ and branch 0 when $Q = 0$. The branch of the returned witness becomes correlated with Q, allowing an observer to infer the hidden bit with better-than-chance accuracy. Thus, relative to an unbiased policy independent of Q, the machine communicates additional information entirely through its legitimate computational choices, without appending a message or sacrificing task performance.

B Task Objective Proofs

B.1 Bounding the Task Objective

We now demonstrate how to bound the task objective. This is achieved through a sequence of steps that focus on different aspects of task performance. For each round t, define the good-estimation

event

$$\mathcal{E}_t = \left\{ |\bar{p}_t - p_t| < \frac{1}{16d} \right\} \cap \left\{ |\hat{q}_t - p_t| < \frac{1}{8d} \right\},$$

where $\bar{p}_t$ is the sample estimate obtained by A prior to adding the perturbation. Let N denote the protocol stopping time. Define the uniform good-estimation event

$$\mathcal{E} = \bigcap_{t=1}^{N} \mathcal{E}_t \cap \{|\hat{p}_N^{\mathcal{O}} - p_N| < \frac{1}{8d}\}$$

and the no-abort event by $\mathcal{K} = \{\text{the protocol does not abort}\}$. So that $\mathcal{E}$ is well-defined, we set $\hat{p}_T^{\mathcal{O}} = p_T$ if no abort occurs. The good-estimation event states allows us to analyse the protocol under well-behaved sample estimates. We first focus on the good-estimation event under small perturbations Δ, which implies that the debate does not abort. That is, if $\hat{q}_t$ and $\hat{p}_t^{\mathcal{O}}$ are "good" estimates, as defined above, then an abort under honest strategy σ_B^{hon} implies

$$|\hat{p}_t - \hat{q}_t| \geq \frac{1}{2d} \implies |\hat{p}_t - p_t| \geq \frac{3}{8d} \implies |\hat{p}_t^{\mathcal{O}} - \hat{p}_t| \geq \frac{1}{4d}, \tag{2}$$

where the last statement contradicts $\mathcal{E}$ for perturbation $\Delta \leq 1/(16d)$. Therefore, an abort is not possible under event $\mathcal{E}$. To bound the task objective on small perturbations, we decompose V's acceptance probability as follows.

$$\Pr[V \to 1 \mid \Delta] = \Pr[V \to 1 \mid \mathcal{E}, \Delta] \Pr[\mathcal{E} \mid \Delta] + \Pr[V \to 1 \mid \neg\mathcal{E}, \Delta] \Pr[\neg\mathcal{E} \mid \Delta]$$

Thus, the task objective satisfies,

$$\Pr[V \to 1 \mid \mathcal{E}, \Delta] \Pr[\mathcal{E}] \leq \Pr[V \to 1, \Delta] \leq \Pr[V \to 1 \mid \mathcal{E}, \Delta] + \Pr[\neg\mathcal{E}], \tag{3}$$

where, for perturbation size $\Delta < 1/(16d)$, $\{\mathcal{E}, \Delta\} \subseteq \mathcal{K}$.

For larger perturbations, we focus on the outcome of an abort under good estimation. By (2), an abort under good estimation implies that $\hat{p}_t$ exceeds the verifier's rejection threshold (which can occur for large enough Δ). Thus, acceptance is impossible on $\mathcal{E} \cap \neg\mathcal{K}$, and hence

$$\Pr[V \to 1, \mathcal{E}, \neg\mathcal{K}] = 0.$$

We conclude that acceptance can occur only on $\mathcal{E} \cap \mathcal{K}$ or on the failure event $\neg\mathcal{E}$. Therefore,

$$\begin{aligned} \Pr[V \to 1] &= \Pr[V \to 1, \mathcal{E}, \mathcal{K}] + \Pr[V \to 1, \mathcal{E}, \neg\mathcal{K}] + \Pr[V \to 1, \neg\mathcal{E}] \\ &\leq \Pr[\mathcal{E} \cap \mathcal{K}] + \Pr[\neg\mathcal{E}]. \end{aligned}$$

If $\Pr[\mathcal{E}] > 0$, this can equivalently be written as

$$\begin{aligned} \Pr[V \to 1] &\leq \Pr[\mathcal{E} \cap \mathcal{K}] + \Pr[\neg\mathcal{E}] \\ &= \Pr[\mathcal{E}] \Pr[\mathcal{K} \mid \mathcal{E}] + \Pr[\neg\mathcal{E}] \\ &= \Pr[\mathcal{E}] \Pr[\mathcal{K} \mid \mathcal{E}] + \Pr[\neg\mathcal{E}] \\ &\leq \Pr[\mathcal{K} \mid \mathcal{E}] + \Pr[\neg\mathcal{E}]. \end{aligned} \tag{4}$$

The following subsections will bound each term in (3) and (4) to provide a characterization of the expected task objective as Δ grows. We begin with the good-estimation event.

B.2 Likelihood of good-estimation

We now provide a bound for the good-estimation event. We require the following version of the Chernoff bound.

Lemma 2. *Let $X_1, \ldots, X_R$ be independent Bernoulli random variables, each taking value 1 with probability p and 0 with probability $(1-p)$. Let $\bar{\mu} = \frac{1}{R}\sum_{i=1}^{R} X_i$ be the empirical mean of the random variables. Then*

$$\Pr\left[|\bar{\mu} - p| \geq s < 2\exp\left(\frac{-s^2 R}{3}\right)\right]$$

Lemma 3. *Suppose that $R \geq 192d^2 \log(100T)$. Then every strategy $\sigma = \sigma_A \circ \sigma_B$ parameterized by $\sigma_A \sim (R, \square)$ and $\sigma_B \sim (R, \square)$ satisfies*

$$\Pr[\neg\mathcal{E}] \leq \frac{1}{300}$$

Proof. If the estimates $\bar{p}_t$, $\hat{g}_t$, and $\hat{p}_t^{\mathcal{O}}$ are the means of R independent Bernoulli samples with probability p_t, Lemma 2 and a union bound gives

$$\Pr\left[|\bar{p}_t - p_t| \geq \frac{1}{16d} \wedge |\hat{q}_t - p_t| \geq \frac{1}{8d}\right] \leq 2\exp\left(-\frac{R}{192d^2}\right) + 2\exp\left(-\frac{R}{96d^2}\right).$$

Therefore, applying a union bound over all T rounds yields

$$\Pr[\neg\mathcal{E}] \leq 4T\exp\left(-\frac{R}{192d^2}\right) \leq \frac{T}{600T} = \frac{1}{600},$$

by the choice of R. Lastly, for V with $r = 192d^2 \log 100$ samples,

$$\Pr\left[|\hat{p}_t^{\mathcal{O}} - p_t| \geq \frac{1}{8d}\right] \leq 2\exp\left(-\frac{r}{48d^2}\right) \leq \frac{4}{100^2} \leq \frac{1}{600}$$

Thus, combining, we observe

$$\begin{aligned}\Pr[\neg\mathcal{E}] &\leq \Pr\left[|\bar{p}_t - p_t| \geq \frac{1}{4d} \wedge |\hat{q}_t - p_t| \geq \frac{1}{4d}\right] + \Pr\left[|\hat{p}_t^{\mathcal{O}} - p_t| \geq \frac{1}{4d}\right] \\ &\leq \frac{1}{600} + \frac{1}{600} = \frac{1}{300}.\end{aligned}$$

□

B.3 Stability of Task Success

We now establish bounds on the task objective under the good estimation event. We record a direct consequence of the Lipschitz property of the oracle machine.

Fact 1. *Let M be K-Lipschitz oracle machine and denote $\Pr[M^O(x) = 1] = \alpha^\star$. Suppose that any non-aborting simulation of M which, at every history $\mathbf{h}$, replaces the reference transition probability $p(\mathbf{h})$ by a possibly randomized, history-dependent probability $\tilde{p}(\mathbf{h})$ satisfying*

$$|\tilde{p}(\mathbf{h}) - p(\mathbf{h})| \leq \frac{1}{d}$$

has acceptance probability within K/d of $\alpha^\star$.

Lemma 4. *Let M be K-Lipschitz oracle machine and denote $\Pr[M^O(x) = 1] = \alpha^\star$. Fix a strategy σ_A with perturbation magnitude $\Delta \leq \frac{3}{4d}$. Then, on the protocol of Figure 2,*

$$\Pr[V \to 1] \leq \alpha^\star + \frac{L}{d} + \frac{1}{300}.$$

If, in addition, for value $\delta \geq 0$, $\{\mathcal{E}, \Delta = \delta\}$ implies that the protocol does not abort, then

$$|\Pr[V \to 1 \mid \mathcal{E}, \Delta = \delta] - \alpha^\star| \leq \frac{L}{d} + \frac{1}{300}.$$

Proof. Construct a simulation $\mathcal{S}$ of the protocol of Figure 2 that follows the reporting strategy σ_A but ignores all abort instructions and continues to the terminal machine output. Extend σ_A to $\mathcal{S}$ on histories that would occur only after an abort.

At each round, let $\hat{p}_t$ be the report proposed by this simulation. Instead of using $\hat{p}_t$ directly, use its projection $\tilde{p}_t$ onto the interval

$$\left[\max\left\{0, p_t - \frac{1}{d}\right\}, \min\left\{1, p_t + \frac{1}{d}\right\}\right].$$

By construction,

$$|\tilde{p}_t - p_t| \le \frac{1}{d}$$

at every history, including histories on which estimation fails. Consequently, if $\tilde{X}$ denotes the terminal output of $\mathcal{S}$, Fact 1 gives

$$\left|\Pr[\tilde{X} = 1] - \alpha^\star\right| \le \frac{K}{d}. \tag{5}$$

Couple $\mathcal{S}$ and the actual protocol using the same random samples and transition randomness for as long as their executions coincide. On $\mathcal{E}$, the perturbation bound implies

$$|\hat{p}_t - p_t| \le |\hat{p}_t - \bar{p}_t| + |\bar{p}_t - p_t| < \frac{3}{4d} + \frac{1}{4d} = \frac{1}{d}.$$

Thus, the projection does not change any report along the actual execution on $\mathcal{E}$. The two simulations therefore coincide until the protocol aborts or terminates.

On $\mathcal{E}$, if the protocol accepts, it cannot have aborted, since, by (2), every abort on $\mathcal{E}$ results in rejection. Its entire computation consequently agrees with the auxiliary simulation, which also outputs 1. Hence,

$$\{V \to 1\} \cap \mathcal{E} \subseteq \{\tilde{X} = 1\}.$$

Therefore,

$$\begin{aligned}\Pr[V \to 1] &\le \Pr[V \to 1, \mathcal{E}] + \Pr[\neg\mathcal{E}] \\ &\le \Pr[\tilde{X} = 1] + \frac{1}{300} \\ &\le \alpha^\star + \frac{K}{d} + \frac{1}{300}.\end{aligned}$$

If $\{\mathcal{E}, \Delta\}$ also implies survival, then the two outputs agree throughout $\mathcal{E}$. Then, for $\delta \ge 0$, where $\{\mathcal{E}, \Delta = \delta\} \subseteq \mathcal{K}$,

$$\Pr[V \to 1 \mid \mathcal{E}, \Delta = \delta] = \Pr[\tilde{X} = 1 \mid \mathcal{E}, \Delta = \delta],$$

and hence

$$\left|\Pr[V \to 1 \mid \mathcal{E}, \Delta = \delta] - \Pr[\tilde{X} = 1]\right| \le \Pr[\neg\mathcal{E}]$$

Combining this inequality with (5) gives

$$|\Pr[V \to 1 \mid \mathcal{E}, \Delta = \delta] - \alpha^\star| \le \frac{L}{d} + \frac{1}{300},$$

as claimed. □

B.4 Non-abort survival

We move to bounding the task objective for general perturbations. By (4), it suffices to provide an upper bound on $\Pr[\mathcal{K} \mid \mathcal{E}]$. We will require the following result. Let $W_1, \ldots, W_n$ be independent, mean-zero random variables with finite third absolute moments and F_{std} denote the cumulative distribution function of the Normal distribution. Writing

$$S_n = \sum_{i=1}^{n} W_i, \qquad \sigma^2 = \sum_{i=1}^{n} \mathbb{E}[W_i^2] > 0, \qquad \rho = \frac{\sum_{i=1}^{n} \mathbb{E}|W_i|^3}{\sigma^3},$$

the Berry-Esseen theorem states the following.

Lemma 5. *For constant $C_{\mathsf{BE}} = 0.56$,*

$$\sup_{x \in \mathbb{R}} \left|\Pr[S_n \le x] - F_{\mathsf{std}}\left(\frac{x}{\sigma}\right)\right| \le C_{\mathsf{BE}} \cdot \rho.$$

We now establish some useful notation. Fix a perturbation magnitude $\Delta \geq 0$. All probabilities in this section refer to this fixed perturbation. Recall that the report at round t is

$$\hat{p}_t = \mathsf{clip}(\bar{p}_t + Q\Delta),$$

and the cross-examiner aborts when

$$|\hat{p}_t - \hat{q}_t| \geq \tau = \frac{1}{2d}.$$

Let $\mathcal{K}_t$ denote the event that no cross-examination abort occurs during the first t rounds, with $\mathcal{K}_0 = \Omega$, and write $\mathcal{K} = \mathcal{K}_T$. Define the estimation-noise difference $Z_t = \bar{p}_t - \hat{q}_t$. In the absence of clipping, the discrepancy equals $Q\Delta + Z_t$. With clipping, however, this identity need not hold. Throughout, assume that $\Pr[\mathcal{E}] > 0$.

B.4.1 One-Round Gaussian Approximation

We first condition only on the past, rather than on the global event $\mathcal{E}$. Throughout we assume that $p_t \in (0.01, 0.99)$. Under this assumption, the event $\mathcal{E}$, and $\Delta < 1/d$, no clipping occurs under the choice of $d = \lceil 150K \rceil$ and $K \geq 1$. For any surviving history $\boldsymbol{h}$, define

$$r_t(\Delta, \boldsymbol{h}) = \Pr[|\hat{p}_t - \hat{q}_t| < \tau \,|\, \mathcal{K}_{t-1}, \mathbf{h}_{t-1} = \boldsymbol{h}].$$

We assume that, conditional on this history and either value of Q having positive conditional probability, the fresh samples

$$X_{t,1}, \ldots, X_{t,R}, Y_{t,1}, \ldots, Y_{t,R}$$

are mutually independent Bernoulli random variables with common mean $p_t = p_t(\boldsymbol{h})$, and that

$$\bar{p}_t = \frac{1}{R}\sum_{i=1}^{R} X_{t,i}, \qquad \hat{q}_t = \frac{1}{R}\sum_{i=1}^{R} Y_{t,i}.$$

In particular, at a fixed history, this fresh-sample law is the same for either value of Q. Define the conditional probability of survival as

$$r_t^0(\Delta, \boldsymbol{h}) = \Pr[|Q\Delta + Z_t| < \tau \,|\, \mathcal{K}_{t-1}, \mathbf{h}_{t-1} = \boldsymbol{h}].$$

For $p_t \in (0.01, 0.99)$, define $\sigma_t = \sqrt{2p_t(1-p_t)/R}$ and

$$g_t(\Delta) = \Pr[|\Delta + \sigma_t z| < \tau],$$

where $z \sim \mathsf{N}(0,1)$.

Lemma 6. *Assume $p_t \in (0.01, 0.99)$ and that no clipping occurs for all configurations of machine input M, then*

$$\left|r_t^0(\Delta, \boldsymbol{h}) - g_t(\Delta)\right| \leq 1/300.$$

Proof. Fix a surviving history $\boldsymbol{h}$ and a value of Q with positive conditional probability. Under this conditioning,

$$Z_t = \frac{1}{R}\sum_{i=1}^{R}(X_{t,i} - p_t) - \frac{1}{R}\sum_{i=1}^{R}(Y_{t,i} - p_t).$$

The summands are independent and centered, and

$$\mathrm{Var}(Z_t) = \frac{2p_t(1-p_t)}{R} = \sigma_t^2.$$

For a Bernoulli random variable X with mean p_t,

$$\begin{aligned}\mathbb{E}[|X - p_t|^3] &= p_t(1-p_t)^3 + (1-p_t)p_t^3 \\ &= p_t(1-p_t)\left[p_t^2 + (1-p_t)^2\right].\end{aligned}$$

Therefore, the standardized sum of third absolute moments is

$$\begin{aligned}\rho_{t,R} &= \frac{2R \cdot R^{-3} \cdot p_t(1-p_t)\left[p_t^2 + (1-p_t)^2\right]}{\sigma_t^3} \\ &= \frac{1}{\sqrt{2R}} \frac{p_t^2 + (1-p_t)^2}{\sqrt{p_t(1-p_t)}}.\end{aligned}$$

Lemma 5 consequently gives, uniformly over x,

$$\left|\Pr[Z_t \le x \,|\, \mathcal{K}_{t-1}, \mathbf{h}_{t-1} = \boldsymbol{h}, Q] - F_{\mathsf{std}}\left(\frac{x}{\sigma_t}\right)\right| \le \frac{C_{\mathsf{BE}}}{\sqrt{2R}} \frac{p_t^2 + (1-p_t)^2}{\sqrt{p_t(1-p_t)}} \le \frac{1}{600},$$

by the choice of R and assumption on p_t.

For $Q = +1$, the unclipped survival event is

$$-\tau - \Delta < Z_t < \tau - \Delta.$$

Applying the distribution-function bound at the interval endpoints yields an error of at most $2/600$ relative to $g_t(\Delta)$. The conditional law of Z_t is symmetric about zero, since it is the difference of two independent, identically distributed sample means. Thus the same interval probability and bound hold for $Q = -1$. Averaging over the conditional distribution of Q proves the result. □

Now we can provide the final conditional survival bound.

Lemma 7. *Fix a perturbation Δ and a round $t \in \{1, \dots, T\}$. Then*

$$\Pr[\mathcal{K} \mid \mathcal{E}] \le \min\left\{1,\, \mathbb{E}[g_t(\Delta) \,|\, \mathcal{K}_{t-1}] + \frac{1}{100}\right\}.$$

Proof. By Lemma 6, for every surviving history $\boldsymbol{h}$, the one-round survival probability satisfies

$$\Pr[\mathcal{K}_t \,|\, \mathcal{K}_{t-1}] \le g_t(\Delta) + \frac{1}{300}.$$

Define $G_t(\Delta) = \mathbb{E}[g_t(\Delta) \,|\, \mathcal{K}_{t-1}]$. Averaging the one-round bound over histories conditional on $\mathcal{K}_{t-1}$ gives

$$\begin{aligned}\Pr[\mathcal{K}_t \mid \mathcal{K}_{t-1}] &= \mathbb{E}[\Pr[\mathcal{K}_t \,|\, \mathcal{K}_{t-1}, \mathbf{h}_{t-1}] \,|\, \mathcal{K}_{t-1}] \\ &\le \mathbb{E}\left[g_t(\Delta) + \frac{1}{300} \,\middle|\, \mathcal{K}_{t-1}\right] \\ &= G_t(\Delta) + \frac{1}{300}.\end{aligned}$$

Since $\mathcal{K} \subseteq \mathcal{K}_t \subseteq \mathcal{K}_{t-1}$, it follows that

$$\begin{aligned}\Pr[\mathcal{K}] &\le \Pr[\mathcal{K}_t] \\ &= \Pr[\mathcal{K}_{t-1}] \Pr[\mathcal{K}_t \mid \mathcal{K}_{t-1}] \\ &\le G_t(\Delta) + \frac{1}{300}.\end{aligned}$$

The law of total probability yields

$$\begin{aligned}\Pr[\mathcal{K} \mid \mathcal{E}] - \Pr[\mathcal{K}] &= \Pr[\neg\mathcal{E}]\,(\Pr[\mathcal{K} \mid \mathcal{E}] - \Pr[\mathcal{K} \mid \neg\mathcal{E}]) \\ &\le \Pr[\neg\mathcal{E}] \\ &\le \frac{1}{150},\end{aligned}$$

By Lemma 3. Consequently,

$$\Pr[\mathcal{K} \mid \mathcal{E}] \le G_t(\Delta) + \frac{1}{300} + \frac{1}{150} = G_t(\Delta) + \frac{1}{100}.$$

Combining this with the trivial bound $\Pr[\mathcal{K} \mid \mathcal{E}] \le 1$ proves the claim. □

B.5 Proof Statements

Theorem 1 follows from (3) with the application of Lemmas 3 and 4 and the observations that $K/d \le 1/150$ and $\{\mathcal{E}, \Delta < 1/(16d)\} \subseteq \mathcal{K}$.

Theorem 2 follows from (4) and the application of Lemmas 3 and 7 for $t = 1$.

C Proof of Theorem 3

We begin with setting notation. Fix $X = x$ and let $\mathbf{y} = (\mathrm{y}_1, \ldots, \mathrm{y}_T)$ be the random variable representing the transcript of $M(x)$ and $\boldsymbol{a} = (a_1, \ldots, a_T)$ denote the realized transcript. Unless stated otherwise, all probabilities, expectations, and information quantities are conditional on $X = x$. Write $h_x = H(Q \mid X = x)$. The view of V is defined as

$$\Pi_V(\mathbf{h}(X, \sigma)) = \begin{cases} (a_1, \ldots, a_T), & \text{on event } \mathcal{K}, \\ (a_1, \ldots, a_t, \hat{p}_t, \{\mathcal{O}(\boldsymbol{a}_{I(t)})\}), & \text{on event } \neg\mathcal{K}_t \mid \mathcal{K}_{t-1}. \end{cases}$$

These transcript types are distinguishable, so the abort indicator and, upon abort, its round are recoverable from Π_V. Notably, V does not observe the internal samples of A or B directly. For some $0 < \beta < 1/2$ we assume that $\Pr[Q = 1 \mid X = x] \in [\beta, 1 - \beta]$. Thus, there is initial uncertainty in the hidden hypothesis.

C.1 Belief updating and confidence truncation

Define the hidden hypothesis posterior as

$$\mu_t = \Pr(Q = 1 \mid \mathbf{h}_{t-1}, \mathcal{K}_{t-1}).$$

These are exact Bayesian posteriors for the observed non-abort channel. At each entering history $\mathbf{h}_{t-1}$, the expected posterior divergence satisfies the identity

$$\mathbb{E}\left[D_{\mathsf{KL}}\big(\mathsf{Bern}(\mu_t) \| \mathsf{Bern}(\mu_{t-1})\big) \,\middle|\, \mathbf{h}_{t-1}, \mathcal{K}_{t-1}\right] = I(Q; \mathrm{y}_t \mid \mathbf{h}_{t-1}, \mathcal{K}_{t-1}). \tag{6}$$

Thus, the expected information gained in an update cannot exceed the remaining uncertainty. More generally, any valid upper bound on the channel's mutual information also bounds the expected posterior divergence in (6). This identity follows from Bayes' rule and is consistent with sequential hypothesis testing, where posterior log-odds accumulate observation log-likelihood ratios.

In our latent disclosure accounting, expected KL increments accumulate according to the mutual-information chain rule and their total is bounded by the initial entropy h_x. We assume that all information used in updating is included in the observed history.

C.2 Single round disclosure

We begin by bounding the amount of information disclosed in a single round of debate. We look at the instances of abort and non-abort separately. For probability laws P and Q, with divergence measured in bits, Pinsker's inequality states that

$$D_{\mathsf{KL}}(P \| Q) \geq \frac{2}{\ln 2} \|P - Q\|_{\mathsf{TV}}^2. \tag{7}$$

Lemma 8. *For $\mu_t \in (\beta, 1 - \beta)$,*

$$I(Q; \mathrm{y}_t \mid \mathbf{h}_{t-1}, \mathcal{K}_{t-1}) \geq \frac{2}{\ln 2} \mu_t (1 - \mu_t) \Delta^2 \geq \frac{2\beta(1 - \beta)}{\ln 2} \Delta^2.$$

Proof. Define the hypothesis conditional sampling probabilities as

$$p_\pm = \Pr[\mathrm{y}_t \mid Q \pm 1, \mathbf{h}_{t-1}, \mathcal{K}_{t-1}]$$

and the mixture $p = \mu_t p_+ + (1 - \mu_t) p_-$. Since a mixture of Bernoulli laws is Bernoulli, the conditional mutual information is

$$I(Q; \mathrm{y}_t \mid \mathbf{h}_{t-1}, \mathcal{K}_{t-1}) = \mu_t D_{\mathsf{KL}}\left(\mathsf{Bern}(p_+) \| \mathsf{Bern}(\bar{p})\right) + (1 - \mu_t) D_{\mathsf{KL}}\left(\mathsf{Bern}(p_-) \| \mathsf{Bern}(\bar{p})\right).$$

For Bernoulli laws, total variation distance equals the absolute difference of their parameters. Thus, (7) gives

$$\begin{aligned} I(Q; \mathrm{y}_t \mid \mathbf{h}_{t-1}, \mathcal{K}_{t-1}) &\geq \frac{2}{\ln 2} \left[\mu_t (p_+ - \bar{p})^2 + (1 - \mu_t)(p_- - \bar{p})^2\right] \\ &= \frac{2}{\ln 2} \mu_t (1 - \mu_t)(p_+ - p_-)^2. \end{aligned} \tag{8}$$

Let ν denote the distribution of $\bar{p}_t$. As the distribution of $\bar{p}_t$ is independent of Q,

$$p_+ - p_- = \int_{[0,1]} \left[\mathsf{clip}(s+\Delta) - \mathsf{clip}(s-\Delta)\right], \nu(ds).$$

For every $s \in [0,1]$ and $0 < \Delta \leq \frac{1}{2}$,

$$\mathsf{clip}(s+\Delta) - \mathsf{clip}(s-\Delta) \geq \Delta.$$

Hence $p_+ - p_- \geq \Delta$. Substitution into (8), followed by $\mu(1-\mu) \geq \beta(1-\beta)$, proves the stated result. Averaging over the entering histories gives the conditional mutual-information bound. □

Disclosure during an abort is given by the following.

Lemma 9. *Suppose that $2R\Delta \notin \mathbb{Z}$. Then, the report $\hat{p}_t$ fully determines Q. In other words,*

$$H(Q \mid \hat{p}_t) = 0.$$

Proof. Recall that the public report upon abort is

$$\hat{p}_t = \frac{K_t}{R} + Q\Delta, \qquad K_t \in \{0, \ldots, R\},$$

Under the two hypotheses, the report belongs respectively to the grids

$$\mathcal{G}_+ = \left\{ \frac{k}{R} + \Delta : 0 \leq k \leq R \right\}, \qquad \mathcal{G}_- = \left\{ \frac{k}{R} - \Delta : 0 \leq k \leq R \right\}.$$

If these grids intersect, there exist integers k, ℓ such that

$$\frac{k}{R} + \Delta = \frac{\ell}{R} - \Delta,$$

which implies

$$2R\Delta = \ell - k \in \mathbb{Z}.$$

The assumed nonintegrality therefore makes the grids disjoint. Membership in the observed grid determines Q. It follows that $H(Q \mid \hat{p}_t) = 0$. □

C.3 Full debate disclosure

We are now in a position to proceed with the main proof of information disclosure. Let h_2 denote the binary entropy function.

Theorem 5. *For $\beta \in (0, \mathbf{1}/2)$ and $\mu_0 \in (\beta, 1-\beta)$, let $k_\beta = \frac{2\beta(1-\beta)}{\ln 2}$. Then*

$$I(Q; \Pi_V(\mathbf{h}) \mid X = x) \geq \frac{h_x - h_2(\beta)}{2} \min\left\{1, \frac{k_\beta \Delta^2 T}{1 - h_2(\beta)}\right\}.$$

Proof. Throughout the proof, condition on $X = x$ and measure entropy in bits. Stop observing the protocol at the first of the following events:

1. an abort;
2. the posterior leaves $(\beta, 1-\beta)$;
3. completion of round T.

Denote the resulting stopped transcript by $\mathbf{u} \subseteq \mathbf{y}$. On history $\mathbf{h}$, since $\mathbf{u}$ is determined by $\Pi_V(\mathbf{h})$, the data-processing inequality gives $I(Q; \Pi_V(\mathbf{h}) \mid X = x) \geq I(Q; \mathbf{u} \mid X = x)$. We bound the latter quantity.

Fix t and an active pre-round history $\mathbf{h}_{t-1}$, and let $\mathcal{A}_t = \neg\mathcal{K}_t \mid \mathcal{K}_{t-1}$ indicate an abort in the current round. Let u_t be the single bit appearing at round t.

Let $N \leq T$ be the number of rounds observed in $\mathbf{u}$. Padding the transcript after stopping with a fixed symbol, the conditional chain rule gives

$$\begin{aligned} I(Q;\mathbf{u}) &= \sum_{t=1}^{T} I(Q;\mathbf{u}_t|\mathbf{h}_{t-1},\mathcal{K}_{t-1}) \\ &= \sum_{t=1}^{T} I(Q;\mathbf{y}_t|\mathbf{h}_{t-1},\mathcal{K}_{t-1}) \cdot \Pr[N \geq t] \\ &\geq k_\beta \Delta^2 \sum_{t=1}^{T} \Pr(N \geq t) \\ &= k_\beta \Delta^2 \, \mathbb{E}[N], \end{aligned}$$

where the inequality follows from Lemma 8. Let $\mathcal{U}$ be the event that, after T rounds, the protocol has neither aborted nor reached the confidence threshold, and set $q = \Pr(\mathcal{U})$. On $\mathcal{U}$, all T rounds are observed. Hence $\mathbb{E}[N] \geq Tq$, and therefore

$$I(Q;\mathbf{u} \mid \mathcal{U}) \geq k_\beta \Delta^2 T q. \tag{9}$$

Outside $\mathcal{U}$, by Lemma 9, on abort posterior entropy is zero by full disclosure and on reaching the confidence threshold it is at most $h_2(\beta)$. Averaging over stopped transcripts gives

$$\begin{aligned} H(Q \mid \mathbf{u}) &\leq h_2(\beta)(1-q) + q \\ &= h_2(\beta) + (1 - h_2(\beta))q. \end{aligned}$$

Since

$$I(Q;\mathbf{u}) = h_x - H(Q \mid \mathbf{u}),$$

we obtain

$$I(Q;\mathbf{u}) \geq h_x - h_2(\beta) - (1 - h_2(\beta))q. \tag{10}$$

Bringing everything together, set $D = h_x - h_2(\beta)$, and $c = 1 - h_2(\beta)$. The assumption $\mu_0 \in (\beta, 1-\beta)$ ensures that $D > 0$. Inequalities (9) and (10) imply that

$$I(Q;\mathbf{u}) \geq \max\left\{k_\beta \Delta^2 T q, D - cq\right\}$$

Multiplying the first inequality by c and the second by $k_\beta \Delta^2 T$, then adding, eliminates q. In other words,

$$\begin{aligned} cI(Q;\mathbf{u}) + k_\beta \Delta^2 T I(Q;\mathbf{u}) &\geq c k_\beta \Delta^2 T q + k_\beta \Delta^2 T (D - cq) \\ &= k_\beta \Delta^2 T D. \end{aligned}$$

Thus,

$$\begin{aligned} (c + k_\beta \Delta^2 T) I(Q;\mathbf{u}) &\geq k_\beta \Delta^2 T D \\ \therefore I(Q;\mathbf{u}) &\geq D \cdot \frac{k_\beta \Delta^2}{c + k_\beta \Delta^2} \end{aligned}$$

Finally, for every $u \geq 0$,

$$\frac{u}{1+u} \geq \frac{1}{2} \min\{1, u\}.$$

Taking

$$u = \frac{k_\beta \Delta^2 T}{c} = \frac{k_\beta \Delta^2 T}{1 - h_2(\beta)}$$

and applying data processing yields

$$I(Q;\Pi_V(\mathbf{h}) \mid X = x) \geq \frac{h_x - h_2(\beta)}{2} \min\left\{1, \frac{k_\beta \Delta^2 T}{1 - h_2(\beta)}\right\}.$$

□

In conclusion, Theorem 3 is a corollary of Theorem 5 for a fixed choice of β.

Debate protocol with error and drift aborts

All parties have access to $\mathcal{O}$, an input $x \in \{0,1\}^n$, and a K-Lipschitz probabilistic oracle machine M. A claims that $\Pr[M(x)=1] \geq \frac{2}{3}$, and the cross-examiner B disputes this claim. Let $W \in [T]$ be the trailing-window length used by an abort drift challenge.

1. Let $\mathbf{y} = (\mathrm{y}_1, \ldots, \mathrm{y}_T)$ be the random variable representing the transcript of $M(x)$ and set $d = \lceil 150K \rceil$.
2. For debate round $t \in [T]$:

(a)-(d) Identical to the equivalent steps in Figure 2.

(e) B can abort error if (it believes) A has deviated on the most recent update or abort drift if (it believes) A has deviated cumulatively over the most recent W updates and $t \geq W$.

3. If abort error was issued in round t:
   (a) V draws $192d^2 \log 100$ independent samples of y_t, conditioned on $\mathrm{y}_{I(t)} = \boldsymbol{a}_{I(t)}$, and uses their sample mean to obtain an estimate $\hat{p}_t^{\mathcal{O}}$ of $\Pr\left[\mathrm{y}_t = 1 \middle| \mathbf{y}_{I(t)} = \boldsymbol{a}_{I(t)}\right]$
   (b) V checks whether $\left|\hat{p}_t^{\mathcal{O}} - \hat{p}_t\right| \geq \frac{1}{4d}$. If the inequality holds, V outputs 0; otherwise, V outputs 1.
4. If abort drift was issued in round t:
   (a) For each $s \in J = \{t-W+1, \ldots, t\}$, V draws $192d^2 \log(100W)$ independent samples of y_s, conditioned on $\mathrm{y}_{I(s)} = \boldsymbol{a}_{I(s)}$, and uses the sample mean to obtain an estimate $\hat{p}_s^{\mathcal{O}}$ of $\Pr\left[\mathrm{y}_s = 1 \middle| \mathbf{y}_{I(s)} = \boldsymbol{a}_{I(s)}\right]$
   (b) V checks whether $\left|\sum_{s \in J}\left(\hat{p}_s^{\mathcal{O}} - \hat{p}_s\right)\right| \geq \frac{\sqrt{N}}{4d}$. If the inequality holds, V outputs 0; otherwise, V outputs 1.
5. If the protocol was not aborted, then the verifier outputs a_T.

Figure 3: Debate protocol for a stochastic oracle with error and drift aborts.

D Proof of Theorem 4

D.1 Concentration Bounds

We require the conditional form of Hoeffding's Lemma.

Lemma 10 (Hoeffding's Lemma)**.** *Let $\mathcal{G}$ be a sigma-algebra, and let Y be a random variable satisfying $a \leq Y \leq b$ almost surely, where a and b are deterministic constants. If $\mathbb{E}[Y \mid \mathcal{G}] = 0$, then, for every real λ,*

$$\mathbb{E}\left[\exp(\lambda Y)|\mathcal{G}\right] \leq \exp\left(\frac{\lambda^2(b-a)^2}{8}\right)$$

almost surely.

The following bounds will be used to bound "good" sampling behavior during protocol execution.

Lemma 11. *Let $(\mathcal{F}_t)_{t \geq 0}$ be a filtration. For each round t, let $p_t \in [0,1]$ be $\mathcal{F}_{t-1}$-measurable. Conditional on $\mathcal{F}_{t-1}$, suppose $X_{t,1}, \ldots, X_{t,R}$ are independent Bernoulli random variables with common mean p_t. Define the $\mathcal{F}_t$-measurable quantities $\hat{p}_t = R^{-1}\sum_{i=1}^R X_{t,i}$, and $D_t = \hat{p}_t - p_t$. Then, almost surely, the following bounds hold.*

1. *For every real λ, $\mathbb{E}\left[\exp(\lambda D_t)|\mathcal{F}_{t-1}\right] \leq \exp\left(\frac{\lambda^2}{8R}\right)$.*
2. *For every $u > 0$, $\Pr\left[|D_t| \geq u|\mathcal{F}_{t-1}\right] \leq 2\exp(-2Ru^2)$.*
3. *For any deterministic integers $0 \leq a < b$ and every $u > 0$,*

$$\Pr\left[\left|\sum_{s=a+1}^{b} D_s\right| \geq u \middle| \mathcal{F}_a\right] \leq 2\exp\left(-\frac{2Ru^2}{b-a}\right).$$

Proof. Conditional on $\mathcal{F}_{t-1}$, each variable $\frac{X_{t,i}-p_t}{R}$ has mean zero and lies in the interval $[-p_t/R, 1-p_t/R]$, whose length is $1/R$. By Lemma 10,

$$\mathbb{E}\left[\exp\left(\lambda\frac{X_{t,i}-p_t}{R}\right)\middle|\mathcal{F}_{t-1}\right] \le \exp\left(\frac{\lambda^2}{8R^2}\right).$$

Conditional independence of the R samples therefore yields

$$\begin{aligned}\mathbb{E}\left[\exp(\lambda D_t)|\mathcal{F}_{t-1}\right] &= \prod_{i=1}^{R}\mathbb{E}\left[\exp\left(\lambda\frac{X_{t,i}-p_t}{R}\right)\middle|\mathcal{F}_{t-1}\right] \\ &\le \exp\left(\frac{\lambda^2}{8R}\right). \end{aligned} \tag{11}$$

This proves the first assertion. For $\lambda > 0$, the Markov inequality gives

$$\Pr\left[D_t \ge u|\mathcal{F}_{t-1}\right] \le \exp\left(-\lambda u + \frac{\lambda^2}{8R}\right).$$

Choosing $\lambda = 4Ru$ gives the upper-tail bound

$$\Pr\left[D_t \ge u|\mathcal{F}_{t-1}\right] \le \exp(-2Ru^2).$$

Applying the same argument to $-D_t$ and adding the two tail bounds proves the second assertion.

For the third assertion, write

$$S_{a,b} = \sum_{s=a+1}^{b} D_s.$$

Since $S_{a,b-1}$ is $\mathcal{F}_{b-1}$-measurable, (11) implies

$$\begin{aligned}\mathbb{E}\left[\exp(\lambda S_{a,b})|\mathcal{F}_a\right] &= \mathbb{E}\left[\exp(\lambda S_{a,b-1})\mathbb{E}\left[\exp(\lambda D_b)|\mathcal{F}_{b-1}\right]|\mathcal{F}_a\right] \\ &\le \exp\left(\frac{\lambda^2}{8R}\right)\mathbb{E}\left[\exp(\lambda S_{a,b-1})|\mathcal{F}_a\right].\end{aligned}$$

Iterating over the $b-a$ rounds gives

$$\mathbb{E}\left[\exp(\lambda S_{a,b})|\mathcal{F}_a\right] \le \exp\left(\frac{(b-a)\lambda^2}{8R}\right).$$

The Markov inequality at $\lambda = \frac{4Ru}{b-a}$, gives

$$\Pr\left[S_{a,b} \ge u|\mathcal{F}_a\right] \le \exp\left(-\frac{2Ru^2}{b-a}\right).$$

Applying the same argument to $-S_{a,b}$ and adding the two tails completes the proof. □

D.2 Protocol Completeness

Given the honest behaviour by both parties, we assume that the transcript update is a fresh Bernoulli draw with parameter $\hat{p}_t$, and that $p_t = \Pr[\mathrm{y}_t \mid \mathbf{y}_{I(t)=\boldsymbol{\alpha}_{I(s)}}$ is the corresponding true transition probability of M. The protocol uses full windows $J_t = \{t-W+1, \ldots, t\}$, with drift challenges permitted only for $t \ge W$. The verifier's drift statistic must sum the s-indexed discrepancies. That that, for probabilistic Turing machine M, the standard bounded-error guarantee is $\Pr[M^{\mathcal{O}}(x) = 1] \ge \frac{2}{3}$.

Lemma 12 (Task correctness)**.** *Let M be a K-Lipschitz probabilistic Turing machine that decides L $R \ge 192d^2\log(100T)$, For any $x \in L$, any fixed window length $W \in [T]$, honest strategy $\sigma_A \sim (R, 0)$, and arbitrary cross-examiner strategy σ_B,*

$$\Pr[V \to 1] \ge \frac{3}{5}.$$

Proof. Simulate an auxiliary execution $\mathcal{S}$ that generates all T transcript updates using the reporting strategy σ_A, ignoring any challenge for purposes of continuing the transcript. Let $\mathbf{y}^{\mathsf{s}}$ denote the random transcript of the simulated execution and before each round t, let $\mathbf{h}^{\mathsf{s}}_{t-1}$ contain the complete history on the simulated execution. Conditional on $\mathbf{h}^{\mathsf{s}}_{t-1}$, the honest reporter draws R fresh, independent Bernoulli samples $X_{t,1}, \dots, X_{t,R}$, each with mean p_t, and reports

$$\hat{p}^{\mathsf{s}}_t = \frac{1}{R} \sum_{j=1}^{R} X_{t,j}.$$

The next transcript bit $\mathrm{y}^{\mathsf{s}}_t$ is generated using independent randomness with success probability $\hat{p}^{\mathsf{s}}_t$. Consequently,

$$\Pr[\mathrm{y}^{\mathsf{s}}_t = 1 \mid \mathbf{h}^{\mathsf{s}}_{t-1}] = \mathbb{E}[\hat{p}^{\mathsf{s}}_t \mid \mathbf{h}^{\mathsf{s}}_{t-1}] = p_t.$$

Thus, the simulated transcript has exactly the transition probabilities of the true execution of $M^{\mathcal{O}}(x)$. If $\mathrm{y}^{\mathsf{s}}_T$ denotes its terminal output, then

$$\Pr[\mathrm{y}^{\mathsf{s}}_T = 1] = \Pr[M^{\mathcal{O}}(x) = 1] \geq \frac{2}{3}.$$

Couple the actual protocol to this simulated execution using the same samples and transcript-update randomness until the actual protocol stops. If no challenge occurs, the actual output equals $\mathrm{y}^{\mathsf{sim}}_T = a^{\mathsf{sim}}_T$. The simulated continuation after a challenge is used only for analysis and is generated independently of the subsequent audit randomness of V.

Define the reporting error $D^{\mathsf{sim}}_t = \hat{p}^{\mathsf{sim}}_t - p_t$. For each round t, let $\mathcal{E}^{(1)}_t$ be the event $|D^{\mathsf{sim}}_t| < \frac{1}{8d}$. For each $t \geq W$, let $\mathcal{E}^{(2)}_t$ be the event

$$\left| \sum_{s \in J_t} D^{\mathsf{sim}}_s \right| < \frac{\sqrt{W}}{8d}.$$

Let $\mathcal{G}^{\mathsf{sim}}$ be the intersection of all these individual and window events in the simulated execution.

By Lemma 11, for every round t,

$$\Pr[\neg \mathcal{E}^{(1)}_t \mid \mathbf{h}^{\mathsf{sim}}_{t-1}] = \Pr\left[|D^{\mathsf{sim}}_t| \geq \frac{1}{8d} \middle| \mathbf{h}^{\mathsf{sim}}_{t-1} \right] \leq 2 \exp\left(-\frac{R}{32d^2} \right).$$

Similarly, for window errors, Lemma 11 gives, for every $u > 0$,

$$\Pr\left[\left| \sum_{s \in J_t} D^{\mathsf{sim}}_s \right| \geq u \mid \mathbf{h}^{\mathsf{sim}}_{t-W} \right] \leq 2 \exp\left(-\frac{2Ru^2}{W} \right).$$

Therefore, it follows that

$$\Pr[\neg \mathcal{E}^{(2)}_t \mid \mathbf{h}^{\mathsf{sim}}_{t-W}] \leq 2 \exp\left(-\frac{R}{32d^2} \right).$$

There are T individual events and at most T window events. Therefore, by a union bound,

$$\begin{aligned} \Pr[\neg \mathcal{G}^{\mathsf{sim}}] &\leq 4T \exp\left(-\frac{R}{32d^2} \right) \\ &\leq 4T(100T)^{-6}. \end{aligned} \tag{12}$$

Let $\mathcal{A}^{(\mathsf{e})}_t$ denote the event that abort error occurs at time t. Condition on a complete pre-audit history $\mathbf{h}^{\mathsf{sim}}$ at which $\mathcal{A}^{(\mathsf{e})}_t$ occurs. This fixes p_t and $\hat{p}_t$. Suppose $\mathcal{G}^{\mathsf{sim}}$ occurs. If V rejects, then

$$|\hat{p}^{\mathcal{O}}_t - \hat{p}_t| \geq \frac{1}{4d},$$

so the triangle inequality implies

$$|\hat{p}^{\mathcal{O}}_t - p_t| > \frac{1}{8d}.$$

V uses $r^{(\mathsf{e})} = \lceil 192d^2 \log 100 \rceil$ fresh, independent samples. Therefore, by Lemma 11,

$$\Pr[V \to 0 \mid \mathbf{h}^{\mathsf{sim}}, \mathcal{G}^{\mathsf{sim}}, \mathcal{A}_t^{(\mathsf{e})}] \le 2\exp\left(-\frac{r^{(\mathsf{e})}}{32d^2}\right) \le 2 \cdot 100^{-6}, \tag{13}$$

under the protocol execution couple to the simulated execution.

Let $\mathcal{A}_t^{(\mathsf{d})}$ denote the event that abort drift occurs at time $t \ge W$. As above, now condition on a complete pre-audit history $\mathbf{h}^{\mathsf{sim}}$ at which $\mathcal{A}_t^{(\mathsf{d})}$ occurs. Suppose $\mathcal{G}^{\mathsf{sim}}$ occurs. For each $s \in J_t$, V draws $r^{(\mathsf{d})} = \lceil 192d^2 \log(100W) \rceil$ fresh samples, independently across all audited rounds and sample indices. Define

$$S_V = \sum_{s \in J_t} (\hat{p}_s^{\mathcal{O}} - p_s).$$

Conditioned on $\mathbf{h}^{\mathsf{sim}}$, S_V is a sum of $Wr^{(\mathsf{d})}$ independent centered variables, each having range length $1/r^{(\mathsf{d})}$. Hence, by Lemma 11

$$\Pr\left[|S_V| \ge u \middle| \mathbf{h}^{\mathsf{sim}}, \mathcal{G}^{\mathsf{sim}}, \mathcal{A}_t^{(\mathsf{d})}\right] \le 2\exp\left(-\frac{2r^{(\mathsf{d})}u^2}{W}\right).$$

If $|S_V| < \sqrt{W}/8d$, the triangle inequality gives

$$\left|\sum_{s \in J_t} (\hat{p}_s^{\mathcal{O}} - \hat{p}_s)\right| < \frac{\sqrt{W}}{4d}.$$

Therefore, V accepts the challenge audit. Consequently,

$$\Pr[V \to 0 \mid \mathbf{h}^{\mathsf{sim}}, \mathcal{G}^{\mathsf{sim}}, \mathcal{A}_t^{(\mathsf{d})}] \le 2\exp\left(-\frac{r^{(\mathsf{d})}}{32d^2}\right) \le 2(100W)^{-6} \le 2 \cdot 100^{-6}. \tag{14}$$

Let $\mathcal{J}$ be the event that a challenge occurs and the verifier outputs 0. By (13) and (14),

$$\begin{aligned}
\Pr[\mathcal{J} \cap \mathcal{G}] &\le \sum_{t,\mathsf{c}} \Pr[\mathcal{A}_t^{(\mathsf{c})} \cap \mathcal{G}^{\mathsf{sim}}] \Pr[V \to 0 \mid \mathcal{A}_t^{(\mathsf{c})} \cap \mathcal{G}^{\mathsf{sim}}] \\
&\le 2 \cdot 100^{-6} \sum_{t,\mathsf{c}} \Pr[\mathcal{A}_t^{(\mathsf{c})} \cap \mathcal{G}^{\mathsf{sim}}] \\
&\le 2 \cdot 100^{-6}.
\end{aligned} \tag{15}$$

Conditional on each pre-audit history, the audit randomness is independent of the auxiliary continuation. Consequently, further conditioning on $\mathcal{G}^{\mathsf{sim}}$ does not change the conditional law of the audit samples.

Under the coupling, if $\mathrm{y}_T^{\mathsf{sim}} = 1$ and the actual protocol outputs 0, then a challenge must have occurred and been rejected. Indeed, without a challenge, the actual output equals $\mathrm{y}_T^{\mathsf{sim}} = 1$. Thus, by (12) and (15),

$$\begin{aligned}
\Pr[V \to 1] &\ge \Pr[\mathrm{y}_T^{\mathsf{sim}} = 1] - \Pr[\mathcal{J}] \\
&\ge \Pr[\mathrm{y}_T^{\mathsf{sim}} = 1] - \Pr[\neg\mathcal{G}^{\mathsf{sim}}] - \Pr[\mathcal{J} \cap \mathcal{G}^{\mathsf{sim}}] \\
&\ge \frac{2}{3} - 4T(100T)^{-6} - 2 \cdot 100^{-6} > \frac{3}{5}
\end{aligned}$$

The argument holds for every legal cross-examiner strategy σ_B, proving the lemma. □

D.3 Smaller Admissible Region

For the second part of Theorem 4 we require the following.

Lemma 13. *Let L be any language decided by a Turing machine M and $W \in [T]$ be the sliding-window parameter. For $R \ge 192d^2 \log(100T)$, $\Delta \ge 1/(d\sqrt{W})$, $\tau_1 = 1/(2d)$, and $\tau_2 = \sqrt{W}/(2d)$, let $\sigma_A \sim (R, \Delta)$, and $\sigma_B \sim (R, \tau_1, \tau_2)$.*

$$\Phi_A(x, \sigma_A \circ \sigma_B) = \Pr[V \leftarrow 1] < \frac{1}{150}.$$

Revised proof. Fix $Q \in \{-1, 1\}$. All probabilities below are conditional on this fixed value of Q. Construct an simulated execution $\mathcal{S}$ that continues for T rounds according to strategies σ_A and σ_B, ignoring challenges. Couple the actual protocol to this simulated execution until its first challenge. The simulated continuation uses randomness independent of the V's fresh audit randomness.

At round t, let p_t be the true oracle probability at the current state $\mathbf{h}^{\mathsf{sim}}$. Let $\bar{p}_t$ and $\hat{q}_t$ be A's and B's unperturbed sample means, each based on R fresh $\mathsf{Bern}(p_t)$ samples. Define $D_A(t)^{\mathsf{sim}} = \bar{p}_t^{\mathsf{sim}} - p_t$, $D_B(t)^{\mathsf{sim}} = \hat{q}_t^{\mathsf{sim}} - p_t$, $S_A(t) = \sum_{s \in J_t} D_A(s)^{\mathsf{sim}}$, and $S_B(t) = \sum_{s \in J_t} D_B(s)^{\mathsf{sim}}$. Note $S_A(t)$ is defined using the unperturbed estimate $\bar{p}_s^{\mathsf{sim}}$, not the perturbed report.

Let $\mathcal{G}_{AB}$ be the event that, for both $i \in \{A, B\}$, $|D_i(t)^{\mathsf{sim}}| < \frac{1}{8d}$for every , for every $t \in \{1, \ldots, T\}$, and $|S_i(t)| < \sqrt{W}/8d$ for every $t \in \{W, \ldots, T\}$

Conditional on the pre-round history $\mathbf{h}_{t-1}^{\mathsf{sim}}$ and Q, each sample mean is unbiased for p_t. By Lemma11, for each $i \in \{A, B\}$,

$$\Pr\left[|D_i(t)^{\mathsf{sim}}| \geq \frac{1}{8d} \middle| Q, \mathbf{h}_{t-1}^{\mathsf{sim}}\right] \leq 2\exp\left(-\frac{R}{32d^2}\right),$$

and

$$\Pr\left[|S_i(t)| \geq \frac{\sqrt{W}}{8d} \middle| Q, \mathbf{h}_{t-W}^{\mathsf{sim}}\right] \leq 2\exp\left(-\frac{R}{32d^2}\right).$$

There are at most $4T$ events in the definition of $\mathcal{G}_{AB}$. Consequently,

$$\begin{aligned}\Pr[\mathcal{G}_{AB}^c \mid Q] &\leq 8T\exp\left(-\frac{R}{32d^2}\right) \\ &\leq 8T(100T)^{-6}. \end{aligned} \tag{16}$$

On $\mathcal{G}_{AB}$, the assumed absence of clipping gives $\hat{p}_t^{\mathsf{sim}} = \bar{p}_t^{\mathsf{sim}} + Q\Delta$. Under the coupled execution. if an abort error challenge has not already occurred, the execution reaches the first full window J_W. At that window, for $\Delta \geq 1/(d\sqrt{W})$,

$$\begin{aligned}\left|\sum_{s \in J_W} (\hat{p}_s - \hat{q}_s)\right| &= |QW\Delta + S_A(W) - S_B(W)| \\ &\geq W\Delta - |S_A(W)| - |S_B(W)| \\ &> \frac{\sqrt{W}}{d} - \frac{2\sqrt{W}}{8d} \\ &= \frac{3\sqrt{W}}{4d} \\ &> \frac{\sqrt{W}}{2d}. \end{aligned} \tag{17}$$

Thus B issues a challenge no later than round W. It remains to show that either possible challenge type leads to output 0, except with uniformly small audit probability.

Let $\mathcal{A}_t^{(\mathsf{e})}$ denote the event that abort error occurs at time t. Condition on a complete pre-audit history $\mathbf{h}^{\mathsf{sim}}$ at which $\mathcal{A}_t^{(\mathsf{e})}$ occurs. This fixes $\bar{p}_t^{\mathsf{sim}}, \hat{q}_t^{\mathsf{sim}}$, and p_t. By the challenge rule,

$$|\hat{p}_t^{\mathsf{sim}} - \hat{q}_t^{\mathsf{sim}}| > \frac{1}{2d}.$$

On events $\mathcal{G}_{AB}$ and $|\hat{q}_t - p_t| < 1/(8d)$ the following holds

$$|\hat{p}_t^{\mathsf{sim}} - p_t| \geq |\hat{p}_t^{\mathsf{sim}} - \hat{q}_t^{\mathsf{sim}}| - |\hat{q}_t^{\mathsf{sim}} - p_t| > \frac{1}{2d} - \frac{1}{8d} = \frac{3}{8d}.$$

If the V's fresh estimate satisfies $|\hat{p}_t^{\mathcal{O}} - p_t| < 1/(8d)$, then

$$\begin{aligned}|\hat{p}_t^{\mathcal{O}} - \hat{p}_t^{\mathsf{sim}}| &\geq |\hat{p}_t^{\mathsf{sim}} - p_t| - |\hat{p}_t^{\mathcal{O}} - p_t| \\ &> \frac{3}{8d} - \frac{1}{8d} \\ &= \frac{1}{4d}. \end{aligned}$$

Hence, V outputs 0. It follows that

$$\Pr[V \to 1 \mid \mathcal{A}_t^{(\mathsf{e})}, \mathbf{h}^{\mathsf{sim}}, \mathcal{G}_{AB}] \le \Pr[|\hat{p}_t^{\mathcal{O}} - p_t| > 1/(8d)] \\ \ge 2\exp\left(-\frac{r^{(\mathsf{e})}}{32d^2}\right) \le 2 \cdot 100^{-6}. \tag{18}$$

Similarly, let $\mathcal{A}_t^{(\mathsf{d})}$ denote the event that abort drift occurs at time t. Condition on a complete pre-audit history $\mathbf{h}^{\mathsf{sim}}$ at which $\mathcal{A}_t^{(\mathsf{d})}$ occurs. Define

$$S_V(t) = \sum_{s \in J_t} \left(\hat{p}_s^{\mathcal{O}} - p_s\right).$$

On $\mathcal{G}_{AB}$ and the event $|S_V(t)| < \frac{\sqrt{W}}{8d}$,

$$\begin{aligned} \left|\sum_{s \in J_t} (\hat{p}_s - \hat{p}_s^{\mathcal{O}})\right| &= |QW\Delta + S_A(t) - S_V(t)| \\ &\ge W\Delta - |S_A(t)| - |S_V(t)| \\ &> \frac{\sqrt{W}}{d} - \frac{\sqrt{W}}{8d} - \frac{\sqrt{W}}{8d} \\ &= \frac{3\sqrt{W}}{4d} \\ &> \frac{\sqrt{W}}{4d}. \end{aligned}$$

Therefore V again outputs 0 under good sample estimates. Thus, by Lemma 11,

$$\Pr[V \to 1 \mid \mathcal{A}_t^{(\mathsf{d})}, \mathbf{h}^{\mathsf{sim}}, \mathcal{G}_{AB}] \le \Pr\left[|S_V(t)| \ge \frac{\sqrt{W}}{8d} \middle| \mathbf{h}^{\mathsf{sim}}, Q\right] \\ \le 2\exp\left(-\frac{r^{(\mathsf{d})}}{32d^2}\right) \le 2(100W)^{-6} \le 2 \cdot 100^{-6}. \tag{19}$$

By (17), on $\mathcal{G}_{AB}$, a challenge necessarily occurs at a time $t \le W$. Its time and type select one audit. That is, they do not create multiple audit-failure penalties. By (18) and (19),

$$\begin{aligned} \Pr[V \to 1, \mathcal{G}_{AB} | Q] &= \sum_{t=1}^{W} \sum_{\mathsf{c}} \Pr[V \to 1 \cap \mathcal{A}_t^{(\mathsf{c})} \cap \mathcal{G}_{AB} | Q] \\ &= \sum_{t=1}^{W} \sum_{\mathsf{c}} \Pr[V \to 1 \mid \mathcal{A}_t^{(\mathsf{c})} \cap \mathcal{G}_{AB} Q] \cdot \Pr[\mathcal{A}_t^{(\mathsf{c})} \cap \mathcal{G}_{AB} \mid Q] \\ &\le 2 \cdot 100^{-6} \sum_{t=1}^{W} \sum_{\mathsf{c}} \Pr[V \to 1 \Pr[\mathcal{A}_t^{(\mathsf{c})} \cap \mathcal{G}_{AB} \mid Q] \\ &\le 2 \cdot 100^{-6}. \end{aligned}$$

Together with (16),

$$\begin{aligned} \Pr[V \to 1 \mid Q] &\le \Pr[\neg \mathcal{G}_{AB} \mid Q] + \Pr[V \to 1, \mathcal{G}_{AB} \mid Q] \\ &\le 8T(100T)^{-6} + 2 \cdot 100^{-6} \\ &< \frac{1}{150}. \end{aligned}$$

The bound is uniform in Q. Averaging over Q completes the proof. □

E AI USE STATEMENT

We used generative AI tools for generating Figure 1 and detecting grammar and spelling errors. All AI-assisted content was reviewed and verified by the authors. The authors take full responsibility for the final content, theoretical claims, and conclusions of this work.