THEORETICAL POSSIBILITIES AND LIMITS OF EFFICIENT CORRUPTION DETECTION IN MULTI-AGENTS

Anonymous authors
Paper under double-blind review

ABSTRACT

This work initiates a theoretical study of the query and communication complexity of detecting a corrupted matrix via queries in the form of linear probes. This is inspired by corruption detection in multi-agents, since reliable multi-agent systems need to detect changes in worker behavior without exchanging complete models or exhaustive output tables. We establish matching upper and lower bounds for query and total communication bits using deterministic and private randomized protocols, both for verifying a single worker and for identifying a corrupted worker in a complete network. These results confirm the possibility of detection with efficient query and communication. In particular, we show that randomization can yield an exponential improvement over deterministic ones and thus achieve efficient query bounds. Our results provide a theoretical foundation for verification mechanisms in multi-agent systems.

1 INTRODUCTION

As large models are increasingly deployed as autonomous agents in multi-agent systems (Park et al., 2023; Wang et al., 2024; Wu et al., 2024), ensuring the reliability of their outputs remains a critical challenge. Agents can generate factually incorrect, logically inconsistent, or structurally corrupted outputs because of failures or adversarial interference (Yao et al., 2022; Kalai et al., 2025; Suwansathit et al., 2026). A change in one worker's underlying behavior can therefore propagate to the final output. Detecting that change efficiently is a basic requirement for reliable delegation.

To study this problem, we introduce a theoretical abstraction inspired by communication complexity. An agent maps each input to a distribution over outputs. For finite input and output spaces, its behavior is represented by a matrix P, whose entry $P_{o,x}$ is the probability of output o given input x. A probability vector q specifies a random input, and the corresponding output distribution is Pq. Corruption detection then becomes matrix difference detection: a supervisor holds a ground-truth reference matrix A, while worker i holds a matrix B_i. We study both single-worker verification (verifying whether one worker agrees with A) and identification among multiple workers (identifying the unique corrupted worker among many otherwise identical workers). The central question is:

What are the query and communication costs needed to identify the corrupted worker?

This question is consequential for recent agents using large models, in particular, large language models (LLMs): their input and output spaces can grow exponentially with sequence length, making full-table communication prohibitively expensive. Efficient verification must therefore use both few probes and short messages.

We develop this framework in two settings. The *discretized setting* uses bounded integer matrices and queries, representing finite-precision scores or counts. The *probabilistic setting* uses conditional-probability matrices and probability-vector queries but communicates in bits, with a known minimum separation between honest and corrupted behavior. Both settings use exact local linear probes and faithful probing controllers: corruption changes the stored matrix, while the controller carries out the verification protocol as specified. This gives a common foundation for comparing single-worker verification and identification in worker networks.

Our results establish matching worst-case query and communication bounds in both settings. Their high-level implications are as follows.

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.

1. **Randomization gives exponential query savings over deterministic methods.** Deterministic verification requires a number of queries nearly linear in the input-space size, whereas randomized verification removes this dependence. When the input space grows exponentially with sequence length, randomization thus removes an exponential query cost.
2. **Randomization+Hashing makes communication-efficient detection possible.** Few probes alone do not guarantee short messages, because each response may still be high-dimensional. Hashing the responses and transmitting compact indices for privately chosen queries and hash functions reduces communication from the scale of a full behavior table to logarithmic dependence on its description length.
3. **Complete worker networks offer no asymptotic advantage over a star network.** With multiple workers, the optimal query and total communication costs grow linearly with the number of workers. Protocols using only supervisor–worker links already attain the optimal asymptotic orders, even when the competing protocols can use arbitrary worker-to-worker communication.

Related Work on Multi-Agent Reliability. Reflection methods such as Reflexion (Shinn et al., 2023) and Retroformer (Yao et al., 2024) seek to improve an agent's outputs using feedback, while studies of self-correction document the difficulty of doing so without reliable external signals (Huang et al., 2023; Xu et al., 2024). Multi-agent debate distributes checking across models (Du et al., 2024; Liu et al., 2024; Liang et al., 2024). Tool-augmented systems provide another source of external information (Schick et al., 2023; Yao et al., 2022; Hao et al., 2023; Mialon et al., 2023). Recursive architectures such as ReDel use supervisors that delegate and aggregate work (Zhu et al., 2024; Zhang et al., 2025; Yang et al., 2026). Our analysis complements these mechanisms by isolating an information cost: the number of exact behavioral probes and transmitted bits needed to detect a specified matrix difference. It does not model the semantic correctness of an LLM answer or establish empirical guarantees for any of these systems. Recent studies categorize multi-agent failures (Cemri et al., 2025), assess resilience to faulty agents (Huang et al., 2025), and detect anomalous interactions using temporal graphs (Zhou et al., 2025). Our framework complements these empirical approaches by characterizing the query and communication costs of detecting fixed behavioral changes, assuming a known reference and faithful probing controllers.

2 A Theoretical Framework For Corruption Detection: Matrix Difference Detection via Linear Queries

We preserve the input–output view of an agent while making the arithmetic, access model, and charged resources explicit. For a positive integer s, write $[s] = \{1, \dots, s\}$. Throughout, $N_{\text{out}} \geq 1$ is the output space size and $N_{\text{in}} \geq 1$ is the input space size.

Agents as Distribution Tables. An agent's probabilistic behavior is a matrix P with entries

$$P_{o,x} = \Pr[\text{output } o \mid \text{input } x],$$

for $o \in [N_{\text{out}}]$ and $x \in [N_{\text{in}}]$. The inputs and outputs may represent token sequences, structured tool calls, reasoning traces, or final answers. This matrix description separates externally observable behavior from the architecture implementing it.

Queries as Matrix-Vector Probes. Let e_x denote the vector with a one at coordinate x and zeros elsewhere. A probe at input x returns the column Pe_x. More generally, a probe following a distribution is represented as a probability vector p over inputs and produces the distribution Pp. Such a probing formulation is standard in property testing and communication complexity, and it captures an important operational reality of agents we are interested in: regardless of their internal structure, only their input-output behaviors are observed in the detection, without direct inspection of model parameters or hidden states.

Corruption Detection Problem. Suppose a supervisor Alice holds the ground-truth model and would like to check if some worker holds a corrupted model. Formally, Alice holds a private reference matrix A. In single-worker verification, a worker has B; Alice should accept when $B = A$ and

reject otherwise. In multiple-worker identification, there are $t \geq 2$ workers with matrices $B_1, \ldots, B_t$. Exactly one index $i^\star \in [t]$ satisfies $B_{i^\star} \neq A$, and all other matrices equal A; Alice should output $i^\star$.

Each worker has a faithful controller that accesses its matrix only through counted oracle calls $q \mapsto B_i q$. It may compute on these answers and send encoded messages, but it has no free interface that reads the matrix. Alice knows her reference A. (All bounds remain the same if assume Alice also accesses A only via linear probes.) The corruption affects the matrix entries, not the agent's adherence to the protocol. All matrices are fixed before the random coins are drawn, and the changed matrix may differ in any nonempty set of entries.

To formalize the finite-precision compute/communication, the main text considers a setting where the matrix/query entries are integers.

- **Discretized Setting**: The matrices A, B, B_i's are in $\mathcal{X} = \mathcal{F}_M^{N_{\text{out}} \times N_{\text{in}}}$ and query vectors in $\mathcal{F}_M^{N_{\text{in}}}$, where the alphabet $\mathcal{F}_M = \{0, 1, \ldots, M-1\}$ for a fixed integer $M \geq 3$.

This simplification isolates the key information costs and helps illustrate the key ideas. Section 5 and Appendices D–F also analyze normalized probability tables directly. The communication costs are similar to the discretized setting, and thus justifies the simplification and the obtained insights.

- **Probabilistic Setting**: The matrices $A, B, B_i \in \mathbb{R}^{N_{\text{out}} \times N_{\text{in}}}$ and query vectors in $\mathbb{R}^{N_{\text{in}}}$; the matrix columns and queries are probabilistic vectors, i.e., with nonnegative entries that sum up to 1. A known threshold $\nu > 0$ lower-bounds the total-variation change of at least one column. Communication consists of binary messages.

Communication and Randomness Setup. A single worker communicates with Alice. For the multiple-worker results, Alice and all workers form a complete communication network: workers may exchange messages as well as communicate with Alice. We use a token-based convention under which an isolated worker's crossing transcript can be simulated as a two-party conversation; Appendix A gives the precise convention. The constructions can use only Alice–worker routes, but their lower bounds allow arbitrary peer communication.

Private randomness is considered. Private random tapes are independent initially, and transmitting a random seed costs bits. No public random coins are provided. (Results using public randomness come readily from those for private randomness by removing the communication of private random seeds.) Local computation and input-independent preprocessing are unrestricted; exact sampling of finite rational laws may use an almost-surely finite number of private fair bits.

Worst-case Guarantees. Write Q for the total number of queries used and C_{tot} for all communicated bits; both costs are hard upper bounds, rather than expectations. Fix a target error $0 < \varepsilon \leq 1/3$. In verification, both false acceptance and false rejection must be at most ε on each fixed input. A one-sided verifier always accepts $B = A$ and can err only by accepting an unequal matrix. In identification, the probability of a wrong index must be at most ε.

For $t = 1$ (verification) or $t \geq 2$ (identification), let Q_t^{det} and C_t^{det} be the minimum worst-case query and total-bit caps of deterministic correct protocols. Let $Q_t^{\text{rand}}(\varepsilon)$ and $C_t^{\text{rand}}(\varepsilon)$ be the analogous private-coin optima. Each optimum ranges over all allowed references and worker matrices. We focus on large $N_{\text{out}}, N_{\text{in}}$ (capturing the extremely large input/output spaces of LLMs); asymptotic notations O, Θ etc are w.r.t. $N_{\text{out}}, N_{\text{in}} \to \infty$ and hide absolute constants. The discretized randomized theorems impose the explicit sufficient condition $N_{\text{in}} \geq 1/\varepsilon$, allowing general confidence while retaining matching bounds.

3 Results For Single Worker

Two key quantities are used often in the bounds:

$$N = N_{\text{out}} N_{\text{in}} \log_2 M = \log_2 |\mathcal{X}|, \qquad L = \log_2(1/\varepsilon). \tag{1}$$

Theorem 3.1 (Discretized setting, single worker, deterministic verification)**.** *For every $N_{\text{out}}, N_{\text{in}} \geq 1$ and $M \geq 3$, the worst-case optimal resources of deterministic correct protocols for single-worker*

verification are

$$Q_1^{\det} = \Theta\left(\frac{N_{\text{in}} \log M}{\log(N_{\text{in}} M)}\right) \qquad \textit{number of queries}, \tag{2}$$

$$C_1^{\det} = \Theta(N) \qquad \textit{total communication bits}. \tag{3}$$

In this scenario, deterministic verification needs order $N_{\text{in}}/\log N_{\text{in}}$ probes. A probe measures every output row simultaneously, explaining why N_{out} does not enter this query rate. It does enter the optimal communication: in the worst case, exact verification requires as many bits, up to constants, as describing the full matrix.

Proof sketch. Stack m queries as columns in a matrix S. For a ground-truth reference A, define $D(A)$ to be the minimum number of columns in S such that

$$\{e \in \mathcal{F}_M^{N_{\text{in}}} - a : eS = 0\} = \{0\} \quad \text{for every row } a \text{ of } A. \tag{4}$$

Equivalently, $\forall B \in \mathcal{X}$, $BS = AS$ implies $B = A$. The optimal number of queries is exactly $D(A)$.

Query lower bound. For a reference row constantly equal to $a = \lfloor (M-1)/2 \rfloor$, a collision $xS = yS$ between distinct $x, y \in \{0, \dots, a\}^{N_{\text{in}}}$ gives such a corruption by adding $x - y$ to that row. Hence the measured sums must distinguish all these vectors, requiring $(1 + N_{\text{in}} a(M-1))^m \geq (a+1)^{N_{\text{in}}}$. Taking logarithms gives the lower bound.

Matching construction. Consdier a uniform binary query on a row difference with k nonzero entries. After flipping the random bits associated with negative coefficients, a zero response corresponds to choosing a subset whose positive weights sum to a fixed value, so these subsets form an antichain, which can be used to bound the miss probability by $\min\{1/2, 1/\sqrt{k+1}\}$. A probabilistic argument shows the existence of a fixed universal list of U queries where

$$U = \min\left\{ \left\lceil \frac{2N_{\text{in}}}{3} \right\rceil, \left\lceil \frac{16 N_{\text{in}} \log(2M)}{\log(N_{\text{in}} + 1)} \right\rceil \right\}. \tag{5}$$

This means $D(A) \leq U$, giving the upper bound on the query number.

Communication. The worker reconstructs its matrix locally and sends its $\lceil N \rceil$-bit index. Conversely, a deterministic accepting transcript shared by distinct equal pairs (A, A) and (B, B) would also accept the unequal pair (A, B). Thus 2^N equal pairs require distinct transcripts and $\Omega(N)$ bits. Appendix B.1 contains the detailed proof, and the corresponding algorithm. □

The proof of Theorem 3.1 supplies an input-independent list of U queries that reconstructs every legal matrix. The randomized protocol uses at most $m_\varepsilon = \min\{U, \lceil \log_M(8/\varepsilon) \rceil\}$ probes and compresses their answers into a short message.

Theorem 3.2 (Discretized setting, single worker, randomized verification)**.** *Let $M \geq 3$, $N_{\text{out}} \geq 1$, $0 < \varepsilon \leq 1/3$, and $N_{\text{in}} \geq 1/\varepsilon$. For single-worker verification in the discretized setting, the worst-case optimal resources of randomized protocols with error at most ε are*

$$Q_1^{\text{rand}}(\varepsilon) = \Theta(\lceil \log_M(1/\varepsilon) \rceil) \qquad \textit{number of queries}, \tag{6}$$

$$C_1^{\text{rand}}(\varepsilon) = \Theta(\min\{N, \log_2(N+1) + L\}) \qquad \textit{total communication bits}. \tag{7}$$

Algorithm 1 attains both orders simultaneously and has no false rejection.

The result means increasing confidence requires only a logarithmic number of additional probes. This dependence is unavoidable, even when queries adapt to previous answers. A verifier need not transmit the complete output vectors: the worker sends a short hash of its answers, and Alice pays for short indices selecting the queries and hash, requiring only $O(\log_2(N+1) + L)$ communication bits.

Proof sketch. *Query upper bound: A random mixture makes a fixed corruption unlikely to cancel.* Fix a nonzero row of $B - A$ and one of its nonzero coefficients. Conditional on all other query coordinates, at most one value of the remaining coordinate makes the dot product zero. A uniform query in $\mathcal{F}_M^{N_{\text{in}}}$ therefore misses with probability at most $1/M$. With m independent queries the miss probability is at most M^{-m}, and equal matrices always pass. This proves the query upper bound; later we will optimize the messages used to implement it.

Query lower bound: Short query lists necessarily leave many input positions indistinguishable. For a hard reference, use a row with every entry $\lfloor (M-1)/2 \rfloor$ and change a uniformly selected pair of positions by $+1$ and -1. Each input position has a length-m signature: its coefficients across those m queries. The pair corruption is invisible precisely when its two signatures agree, because their contributions cancel. There are at most M^m signatures for N_{in} positions. If n_h counts the positions with signature h, then $\sum_h n_h^2 \geq N_{\text{in}}^2 / M^m$. Writing $[z]_+ = \max\{z, 0\}$, counting pairs in the same cell yields the useful collision bound

$$\Pr\{\text{pair corruption is invisible}\} \geq \varphi(m), \qquad \varphi(u) = \left[\frac{N_{\text{in}} M^{-u} - 1}{N_{\text{in}} - 1} \right]_+ . \tag{8}$$

This leads to the lower bound.

Communication upper bound: Compress the random choice as well as the answer. Sending all random query coordinates is expensive. Instead, construct an input-independent table of query schedules in advance, and sample uniformly from it; a *schedule* means a list of query vectors. For $m = \lceil \log_M(8/\varepsilon) \rceil < U$, concentration and a union bound over fewer than 2^{2N} unequal matrix pairs give a table of $O((N+1)/\varepsilon^2)$ schedules where a uniform sampled schedule has miss probability at most $\varepsilon/4$. Alice transmits its short index, and the worker generates that schedule locally. This table is fixed independently of the inputs; only the chosen index is random and paid for.

Similarly compress the response by a hashing table. The response array has polynomially many bits in N because $m \leq U \leq N_{\text{in}}$ and every entry is at most $N_{\text{in}}(M-1)^2$. An independently selected hash from a fixed finite table compresses it to $O(L)$ bits, using another index of length $O(\log_2(N+1) + L)$. The hash collision probability is at most $\varepsilon/4$.

The two failure events are a missed corruption and a hash collision; independence bounds their sum by $\varepsilon/2$. Note that in the special case $m_\varepsilon = U$, no need to use random schedule index; simply use the universal list of queries from Theorem 3.1. Also, sending the exact matrix instead provides the N-bit alternative. See the branches in Algorithm 1.

Communication lower bound: Too few transcript rectangles can miss unequal pairs. The lower bound again grants full matrix access, reducing to equality on 2^N labels. With b transmitted bits, fixing all random tapes partitions the accepted input pairs into at most 2^{b+1} transcript rectangles, each a product of a set of Alice inputs and a set of worker inputs. Too few rectangles force excessive acceptance of unequal pairs. Averaging over tapes gives $\varepsilon + 2^{-N} \geq c_0 2^{-b}$ for an absolute $c_0 > 0$ and hence the $\min\{N, L\}$ cost. Private random coins impose an additional constraint: each transcript's probability factors into one term for each party. If two inputs have nearly identical vectors of these factors, the protocol cannot reliably distinguish an equal pair from a crossed pair. Packing these vectors forces $b = \Omega(\log_2(N+1))$. Combining the two costs proves the communication minimum.

Appendix B.3 provides the detailed proofs for all the bounds. □

Algorithm 1 presents an algorithm achieving the bounds in Theorem 3.2. For small N, the worker just sends the whole B to Alice for verification. Otherwise, Alice selects m_ε queries and sends them to the worker; the worker probes and returns the response for verification. The tables are used for compressing the sent queries or answers. They are part of the input-independent protocol description (details in Appendix B.3.1).

Algorithm 1 Single-worker, randomized verification (Theorem 3.2)

Input: Alice's reference A, worker matrix B, target error ε, the universal query list from Theorem 3.1, the fixed query and hash tables in Theorem 3.2.
Output: Error at most ε, with no false rejection.

if $N \leq \log_2(N+1) + L$ **then**
  Alice sends a start frame
  Worker reconstructs B with the U universal query list, and sends it via $\lceil N \rceil$-bits.
  return ACCEPT iff that equals Alice's A.
end if
Set $m = m_\varepsilon := \min\{U, \lceil \log_M(8/\varepsilon) \rceil\}$. If $m = U$, use the U universal query list; otherwise Alice selects a m-query schedule by selecting a uniform index in the fixed table.
Alice independently selects a uniform hash-table index and sends all needed indices in a start frame.
Worker makes the m local probes, encodes the responses, and returns its k_ε-bit hash for $k_\varepsilon = \lceil \log_2(16/\varepsilon) \rceil$.
Alice forms her reference response array and returns ACCEPT iff its hash agrees.

4 Results For Multiple Workers

We now allow all peer links and require identification of one corrupted worker among $t \geq 2$ workers. Every local probe and every network bit still counts.

Theorem 4.1 (Discretized setting, multiple workers, deterministic identification in a complete network)**.** *For $N_{\text{out}}, N_{\text{in}} \geq 1$, $M \geq 3$, and $t \geq 2$, for multiple-worker identification in a complete network in the discretized setting, the worst-case optimal resources of deterministic correct protocols are*

$$Q_t^{\text{det}} = \Theta\left(t \frac{N_{\text{in}} \log M}{\log(N_{\text{in}} M)}\right) \qquad \textit{number of queries}, \tag{9}$$

$$C_t^{\text{det}} = \Theta(tN) \qquad \textit{total communication bits}. \tag{10}$$

A deterministic protocol attains both orders simultaneously. The lower bounds allow arbitrary communication between workers and adaptive choice of which worker to probe.

In the worst case, deterministic identification requires enough information to clear essentially every worker, for which the peer communication cannot help. The query complexity is thus in the order of the worker number t times the complexity in the single-worker verification case. Similar for communication: peer communication can change where verification is performed, but cannot reduce the optimal order of total traffic in this model.

Proof sketch. *Queries.* Let $D(A)$ be the minimum size of a separating query list for A. Run the protocol on the all-honest tuple (i.e., all workers are uncorrupted), which must terminate within its hard caps despite lying outside the correctness promise, and call its output g. Every worker $i \neq g$ must receive a separating list: otherwise a legal corruption matching its replies leaves the entire transcript unchanged and is misidentified. This remains true with peer messages because matrix information is only learned through probes. Hence the cap is at least $(t-1)D(A)$; the hard single-worker reference gives the claimed order.

Communication bits. For total cap C, isolate a uniformly chosen worker, give it matrix y, and give Alice and all others x. Accept equality when the output differs from the isolated index. Unequal inputs are never accepted, while equal inputs have mean acceptance $1 - 1/t$. Mean cut traffic is at most $2C/t$, since each peer bit crosses two worker cuts. Truncation at $O(C/t + 1)$ bits, rejecting equality on truncation, retains constant equal-input acceptance by Markov's inequality. Charged framing makes this a valid two-party transcript. The equality lower bound then gives $C = \Omega(tN)$ for growing N.

Simultaneous upper bound. Reconstruct and compare workers $1, \ldots, t-1$ using the universal list from Theorem 3.1 and their transmitted matrix indices; infer the last worker if all match. This uses at most $(t-1)U$ probes and $O(tN)$ total bits.

Appendix C.1 contains the detailed proof, and also gives the corresponding algorithm achieving the upper bounds. □

Theorem 4.2 (Discretized setting, multiple worker, randomized identification in a complete network)**.** *Let $M \geq 3$, $N_{\text{out}} \geq 1$, $t \geq 2$, $0 < \varepsilon \leq 1/3$, and $N_{\text{in}} \geq 1/\varepsilon$. For multiple-worker identification in a complete network in the discretized setting, the worst-case optimal resources of randomized protocols with error at most ε are*

$$Q_t^{\text{rand}}(\varepsilon) = \Theta(t \lceil \log_M(1/\varepsilon) \rceil) \qquad \textit{number of queries}, \tag{11}$$

$$C_t^{\text{rand}}(\varepsilon) = \Theta(t \min\{N, \log_2(N+1) + L\}) \qquad \textit{total communication bits}. \tag{12}$$

Algorithm 2 attains both orders simultaneously.

Each additional worker contributes the same optimal order of probes and bits as a single verification task. Even a complete network cannot improve the order of total queries or total traffic. Also note that peer communication is allowed, while the achieving protocol in Algorithm 2 only uses fixed Alice–worker routes. This means a complete network does improve over a star network. Finally, confidence costs $\lceil \log_M(1/\varepsilon) \rceil$ probes per tested worker, with no extra $\log t$ factor. The reason is that local tests never reject an honest worker: only the unique corrupted worker can cause a missed detection. Thus a target error ε for each local test suffices for the whole network.

Proof sketch. *Upper bounds: Reuse a one-sided test without splitting the error budget.* Test workers $1, \dots, t-1$ with Algorithm 1, stop at the first rejection, and otherwise return t. Every honest worker passes with certainty. If the corrupted worker is tested, only its own missed-detection event can cause an error; if it is last, the fallback identifies it exactly. Each test can therefore use target ε, with no division by t. Summing its hard query and bit caps gives the simultaneous upper bounds. Every transmitted seed copy is charged; the fixed traversal requires only $O(t)$ control bits.

Query lower bound: Use the collision bound to rule out cheaper adaptive allocations. Fix all private tapes and follow the execution. Let g be its fallback output and m_i its number of probes to worker i. Choose a uniformly random corrupted worker and the same coordinate-pair corruption as in the single-worker proof, independently of those tapes. For $i \neq g$, an invisible corruption produces the honest transcript and hence the wrong output. By (8), its conditional miss probability is at least $\varphi(m_i)$. If c is the hard total query cap, convexity and monotonicity give

$$\Pr\{\text{identification error}\} \;\geq\; \frac{1}{t}\sum_{i \neq g} \varphi(m_i) \;\geq\; \frac{t-1}{t}\,\varphi\!\left(\frac{c}{t-1}\right) \quad \text{for each fixed plan.} \tag{13}$$

Here the probability is over the fixed corruption prior. Intuitively, concentrating probes on some workers leaves other workers disproportionately vulnerable. Then requiring error at most ε and using $N_{\text{in}} \geq 1/\varepsilon$ yields $c \geq (t-1)\log_M(1/(3\varepsilon))$. A second bound handles the regime where this expression is small: a c-probe execution path visits at most c workers and can correctly guess only one additional unvisited worker. Thus success against a uniformly located corruption is at most $(c+1)/t$, giving $c \geq t(1-\varepsilon) - 1 = \Omega(t)$. Together the two bounds imply the stated ceiling-logarithmic order.

Communication lower bound: Confidence and private randomness impose separate network costs. The averaged worker-cut reduction above now has false acceptance at most ε. The same truncation and rectangle argument forces $C = \Omega(t \min\{N, L\})$. This confidence argument may expose the coins for analysis, but doing so would lose the additional private-coin cost. To retain that cost, first reduce error to a small absolute constant by a constant number of independent repetitions, then fix each cut separately. Let a_i be the average probability of outputting i on a uniformly chosen all-equal matrix. Since $\sum_i a_i = 1$, at least $t-1$ cuts have $a_i \leq 3/4$, and their induced equality tests accept equal inputs with average probability at least $1/4$.

Let T_i be the mean traffic of such a cut on these all-equal inputs. Truncating at a constant multiple of $T_i + 1$ retains constant acceptance. The two sides still have independent private tapes, so the single-worker private packing bound gives $T_i = \Omega(\log_2(N+1))$ for growing N. Summing these bounds costs at most twice the amplified network traffic. This proves the additional $\Omega(t \log_2(N+1))$ term; combining it with the confidence term gives the claimed minimum. Thus neither adaptive routing nor peer collaboration removes the private communication cost.

Appendix C.2 gives the detailed proofs. □

Algorithm 2 Multiple workers, randomized identification in a complete network (Theorem 4.2)

Input: Alice's A; $t \geq 2$ faithful worker oracles; target ε; the fixed tables for Algorithm 1.
Output: Identification error at most ε on every valid tuple.

1: **for** $i = 1, \dots, t-1$ in a fixed traversal **do**
2: Run Algorithm 1 with worker i and target ε, paying all seed copies and start/return frames.
3: Return i at the first rejection.
4: **end for**
5: **return** t.

5 Results For The Probabilistic Setting

We now return to the motivating model of conditional probabilities; only results for randomized verification are presented here, while deterministic ones are in the appendix. The results illustrate the same conclusions as the discretized setting: randomization substantially reduces the number of probes, short encoded messages suffice, and complete worker connectivity does not improve the optimal communication order. The query rates differ because probability-valued queries allow arbitrarily fine mixtures, while the communication costs remain essentially the same, with dependence on the adaptive quantization for communication w.r.t. the known separation ν.

Theoretical Setup. Every column of A and of a worker matrix B is nonnegative and sums to one. A query q is likewise a nonnegative vector summing to one. Write $\mathrm{TV}(p, p') = \frac{1}{2}\|p - p'\|_1$ for total variation between probability vectors. A supplied threshold $0 < \nu \leq 1/8$ specifies the promise: a worker is either equal to A or has

$$\delta(A, B) := \max_{j \in [N_{\mathrm{in}}]} \mathrm{TV}(A_{\cdot j}, B_{\cdot j}) \geq \nu.$$

In the network there is exactly one such corrupted worker. A known positive threshold is essential: merely assuming $B \neq A$ does not permit a uniform finite private-coin communication bound.

Alice knows her private reference matrix A. All workers learn their matrices only through counted probes, and all transmitted messages are binary. Random choices use private fair bits, finitely many on each terminating execution, with almost-sure termination and hard resource caps on every input. Every seed, frame, and traversed edge is charged. Define $Q^{\mathrm{rand}}_{\mathrm{p},t}(\nu, \varepsilon)$ and $C^{\mathrm{rand}}_{\mathrm{p},t}(\nu, \varepsilon)$ as the separately minimized worst-case query and total-bit caps, respectively, among protocols with both caps finite. As before, $t = 1$ means verification with both errors at most ε, and $t \geq 2$ means identification with error at most ε. Appendix D gives the full details about the formal setup, conventions, and notations.

For the probabilistic setting, the following quantities are used often in the bounds:

$$N_\nu = N_{\mathrm{in}}(N_{\mathrm{out}} - 1)\log_2(1/\nu), \qquad F_\nu(\varepsilon) = \log_2(N_\nu + 1) + L, \qquad L = \log_2(1/\varepsilon). \tag{14}$$

Here N_ν measures the matrix description length at resolution ν (similar to the N in the discretized setting), and $F_\nu(\varepsilon)$ combines the dimension and confidence costs.

Theorem 5.1 (Probabilistic setting, single worker, randomized verification)**.** *Let $N_{\mathrm{out}} \geq 2$, $N_{\mathrm{in}} \geq 1$, $0 < \nu \leq 1/8$, and $0 < \varepsilon \leq 1/3$. For single-worker verification in the probabilistic setting, the optimal worst-case resources of randomized protocols with error at most ε are*

$$Q^{\mathrm{rand}}_{\mathrm{p},1}(\nu, \varepsilon) = 1 \qquad \textit{number of queries}, \tag{15}$$

$$C^{\mathrm{rand}}_{\mathrm{p},1}(\nu, \varepsilon) = \Theta(\min\{N_\nu, F_\nu(\varepsilon)\}) \qquad \textit{total communication bits}. \tag{16}$$

A random mixture detects any fixed corrupted matrix with one probe, whereas deterministic verification can require all N_{in} columns. As in the discretized setting, the worker sends a short hash instead of its full response table, and Alice pays for the private seed. The communication depends only logarithmically on the table's description length at the promised resolution and on $1/\varepsilon$, until transmitting the quantized table becomes cheaper. In the same regime as Theorem 3.2, both resource orders are attained simultaneously. The improved one-query rate relies on exact local probability arithmetic, since probability-valued queries allow arbitrarily fine mixtures.

Proof sketch. *Query bound: Encode a row difference as a polynomial.* For $z \in [1, 2)$ consider the following legal probability query

$$q(z) = \frac{(1, z, \ldots, z^{N_{\mathrm{in}}-1})^\top}{1 + z + \cdots + z^{N_{\mathrm{in}}-1}}.$$

A row of $(B - A)q(z)$ is a polynomial of degree at most $N_{\mathrm{in}} - 1$ divided by the positive denominator. The column-distance promise forces some coefficient in some row to have magnitude at least $2\nu/N_{\mathrm{out}}$. The response discrepancy therefore cannot be uniformly tiny and thus can be detected.

Communication upper bound: Turn exact separation into a finite message. Choose z uniformly from a finite rational grid. Lagrange interpolation shows that at most $N_{\mathrm{in}} - 1$ grid points can have that row's response below a suitable positive threshold in magnitude, since too many small values would force every coefficient to be small. Therefore, a grid of order $N_{\mathrm{in}}/\varepsilon$ makes this bad event unlikely, and quantizing the local response more finely than the threshold then preserves inequality at every good point. Then Alice can send the grid index and an independent finite hash index, and compare the returned hash to her reference hash, which is sufficient for the detection. In this way, only short indices and hash values cross the link, costing $O(F_\nu(\varepsilon))$ bits.

Communication lower bound: Pack matrices. For the lower bound, pack the probability matrices at separation greater than ν. The packing contains $2^{\Omega(N_\nu)}$ matrices, so their verification includes private equality on that many labels. Its bit lower bound matches hashing or, when cheaper, exact quantization of all columns.

Appendix E.2 gives quantitative thresholds, the complete analysis, and a corresponding algorithm. □

Theorem 5.2 (Probabilistic setting, multiple workers, randomized identification in a complete network)**.** *Let $t \geq 2$, $N_{\mathrm{out}} \geq 2$, $N_{\mathrm{in}} \geq 2$, $0 < \nu \leq 1/8$, and $0 < \varepsilon \leq 1/3$. For multiple-worker verification in a complete network in the probabilistic setting, the optimal worst-case resources of randomized protocols with error at most ε are*

$$Q^{\mathrm{rand}}_{\mathrm{p},t}(\nu, \varepsilon) = \lfloor t(1-\varepsilon) \rfloor \qquad \textit{number of queries}, \tag{17}$$

$$C^{\mathrm{rand}}_{\mathrm{p},t}(\nu, \varepsilon) = \Theta(t \min\{N_\nu, F_\nu(\varepsilon)\}) \qquad \textit{total communication bits}. \tag{18}$$

One randomized probe per tested worker is enough. As in Theorem 4.2, there is no extra $\log t$ confidence penalty: an uncorrupted worker always passes the local test, so only the unique corrupted worker can be missed. Peer links do not reduce the optimal order of total traffic. Together, the two theorems show that the main efficiency conclusions persist for normalized conditional probabilities, with the resolution parameter replacing the integer alphabet in the communication bounds in the discretized setting.

Proof sketch. *Upper bounds: Reuse the one-sided verifier.* Testing $t-1$ workers and inferring the last gives one probe per tested worker and $O(tF_\nu(\varepsilon))$ bits. Honest workers never reject, so the error analysis again concerns only the corrupted worker. Selecting exact quantization instead yields the communication saturation branch.

Why the exact query count has an integer boundary. With a cap of c probes, an all-honest execution visits at most c workers. Among unvisited workers, only its fallback can be identified correctly. A uniformly located corruption therefore has success at most $(c+1)/t$. If equality in this coverage bound were attained, every visited worker would have to be recognized without any miss. For $N_{\mathrm{in}} \geq 2$, this is impossible with one probe and finite private coins: some complete finite coin history has positive probability, and, at the reference with every column equal to $(1/2, 1/2, 0, \ldots, 0)^\top$, its single query admits a legal probability-matrix perturbation in its nullspace. That fixed corruption is missed on the whole history. This explains the extra query at the integer boundary. Conversely, test a uniform subset and use a uniform untested fallback, choosing the positive local miss rate small enough to fit the remaining error budget.

Communication lower bounds: Combine worker-cut and packing arguments. To bound the total bits, apply the worker-cut argument to the separated probability packing. Private equality lower bounds sum to $\Omega(t \min\{N_\nu, F_\nu(\varepsilon)\})$, regardless of peer links or additional local probes. Note that the exact query-minimal construction and the bit-optimal construction are distinct.

Appendix F.2 proves both claims and gives a corresponding algorithm. □

6 Conclusions And Limitations

This work formalized corruption detection in multi-agent systems through matrix difference detection. In the discretized setting, matching upper and lower query and total communication bounds were obtained for deterministic and private randomized protocols, both for one worker and for a complete worker network. The randomization reduces the optimal query count to order $\lceil \log_M(1/\varepsilon) \rceil$ per tested worker, and compressing the sent queries/responses via hashing achieves efficient communication. Each case admits a protocol attaining both resource orders simultaneously. In the probabilistic setting, the results support the same efficiency conclusions, with one randomized probe per tested worker under exact local distribution access. Complete peer connectivity does not improve the optimal order of total traffic in either setting.

Limitations. The results assume exact responses for probes, a known reference, one corrupted matrix, faithful controllers, and fully charged communication. They do not address noise, approximate equality, multiple corrupted workers, or Byzantine message fabrication, which are interesting future directions. The matching randomized query formulas for the discretized setting require $N_{\mathrm{in}} \geq 1/\varepsilon$; they are not claimed uniformly when confidence grows faster than the input dimension permits. The four discretized theorems establish simultaneous optimal asymptotic orders, rather than simultaneous exact minimum query and bit counts. The probabilistic results give the same qualitative efficiency conclusions under a supplied separation threshold and exact local distribution responses; their joint optimality conditions are stated separately. No empirical claim about present LLM verification follows without additional modeling and experiments.

AI USE STATEMENT

Generative AI tools assisted with editing assistance (grammar, spelling, and word choice). They also assisted with organizing the manuscript and compilation checks. They were used in interactions with the authors to derive the lower bounds for randomized protocols. They were not used for developing the theoretical formulation, and the mathematical claims for the other results. The authors have reviewed all AI-assisted content, and take responsibility for the final content of this work.

REFERENCES

Mert Cemri, Melissa Z. Pan, Shuyi Yang, Lakshya A. Agrawal, Bhavya Chopra, Rishabh Tiwari, Kurt Keutzer, Aditya Parameswaran, Dan Klein, Kannan Ramchandran, Matei A. Zaharia, Joseph E. Gonzalez, and Ion Stoica. Why do multi-agent LLM systems fail? In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-4082. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/b1041e52d3be19f0a9bc491657488e4a-Abstract-Datasets_and_Benchmarks_Track.html.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B Tenenbaum, and Igor Mordatch. Improving factuality and reasoning in language models through multiagent debate. In *Forty-first international conference on machine learning*, 2024.

Shibo Hao, Tianyang Liu, Zhen Wang, and Zhiting Hu. Toolkengpt: Augmenting frozen language models with massive tools via tool embeddings. *Advances in neural information processing systems*, 36:45870–45894, 2023.

Jen-Tse Huang, Jiaxu Zhou, Tailin Jin, Xuhui Zhou, Zixi Chen, Wenxuan Wang, Youliang Yuan, Michael Lyu, and Maarten Sap. On the resilience of LLM-based multi-agent collaboration with faulty agents. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 26202–26226. PMLR, 2025. URL https://proceedings.mlr.press/v267/huang25ay.html.

Jie Huang, Xinyun Chen, Swaroop Mishra, Huaixiu Steven Zheng, Adams Wei Yu, Xinying Song, and Denny Zhou. Large language models cannot self-correct reasoning yet. *arXiv preprint arXiv:2310.01798*, 2023.

Adam Tauman Kalai, Ofir Nachum, Santosh S Vempala, and Edwin Zhang. Why language models hallucinate. *arXiv preprint arXiv:2509.04664*, 2025.

Tian Liang, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, Shuming Shi, and Zhaopeng Tu. Encouraging divergent thinking in large language models through multi-agent debate. In *Proceedings of the 2024 conference on empirical methods in natural language processing*, pp. 17889–17904, 2024.

Tongxuan Liu, Xingyu Wang, Weizhe Huang, Wenjiang Xu, Yuting Zeng, Lei Jiang, Hailong Yang, and Jing Li. Groupdebate: Enhancing the efficiency of multi-agent debate using group discussion. *arXiv preprint arXiv:2409.14051*, 2024.

Grégoire Mialon, Roberto Dessi, Maria Lomeli, Christoforos Nalmpantis, Ramakanth Pasunuru, Roberta Raileanu, Baptiste Roziere, Timo Schick, Jane Dwivedi-Yu, Asli Celikyilmaz, et al. Augmented language models: a survey. *Transactions on Machine Learning Research*, 2023.

Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th annual acm symposium on user interface software and technology*, pp. 1–22, 2023.

Timo Schick, Jane Dwivedi-Yu, Roberto Dessi, Roberta Raileanu, Maria Lomeli, Eric Hambro, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. Toolformer: Language models can teach themselves to use tools. *Advances in neural information processing systems*, 36:68539–68551, 2023.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning. *Advances in neural information processing systems*, 36:8634–8652, 2023.

Surada Suwansathit, Yuxuan Zhang, and Guofei Gu. A systematic taxonomy of security vulnerabilities in the openclaw ai agent framework. *arXiv preprint arXiv:2603.27517*, 2026.

Guanzhi Wang, Yuqi Xie, Yunfan Jiang, Ajay Mandlekar, Chaowei Xiao, Yuke Zhu, Linxi Fan, and Anima Anandkumar. Voyager: An open-ended embodied agent with large language models. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL https://openreview.net/forum?id=ehfRiF0R3a.

Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Beibin Li, Erkang Zhu, Li Jiang, Xiaoyun Zhang, Shaokun Zhang, Jiale Liu, et al. Autogen: Enabling next-gen llm applications via multi-agent conversations. In *First conference on language modeling*, 2024.

Wenda Xu, Guanglei Zhu, Xuandong Zhao, Liangming Pan, Lei Li, and William Wang. Pride and prejudice: Llm amplifies self-bias in self-refinement. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15474–15492, 2024.

Chenxiao Yang, Nathan Srebro, and Zhiyuan Li. Recursive models for long-horizon reasoning. *arXiv preprint arXiv:2603.02112*, 2026.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. In *The eleventh international conference on learning representations (ICLR)*, 2022.

Weiran Yao, Shelby Heinecke, Juan Carlos Niebles, Zhiwei Liu, Yihao Feng, Le Xue, Rithesh Ramapura Narasimha Murthy, Zeyuan Chen, Jianguo Zhang, Devansh Arpit, et al. Retroformer: Retrospective large language agents with policy gradient optimization. In *International Conference on Learning Representations*, volume 2024, pp. 10091–10111, 2024.

Alex L Zhang, Tim Kraska, and Omar Khattab. Recursive language models. *arXiv preprint arXiv:2512.24601*, 2025.

Jialong Zhou, Lichao Wang, and Xiao Yang. GUARDIAN: Safeguarding LLM multi-agent collaborations with temporal graph modeling. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-0272. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/0bc795afae289ed465a65a3b4b1f4eb7-Abstract-Conference.html.

Andrew Zhu, Liam Dugan, and Chris Callison-Burch. Redel: A toolkit for llm-powered recursive multi-agent systems. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 162–171, 2024.

Appendix

CONTENTS

1 Introduction **1**

2 A Theoretical Framework For Corruption Detection: Matrix Difference Detection via Linear Queries **2**

3 Results For Single Worker **3**

4 Results For Multiple Workers **6**

5 Results For The Probabilistic Setting **7**

6 Conclusions And Limitations **9**

A Discretized Setting: Communication Conventions And Notation **13**

B Discretized Setting: Proofs For Single-Worker Verification **13**

B.1 Deterministic Verification: Proof for Theorem 3.1 13

B.2 Separating Queries Characterization 13

B.2.1 Communication Lower Bounds For Deterministic Potocols For A Single Worker . . . 15

B.3 Randomized Verification: Proof for Theorem 3.2 16

B.3.1 Private-Coin Encoding and Resource Upper Bounds 16

B.3.2 Communication Lower Bounds For A Single Worker 17

B.3.3 Exact Randomized Query Complexity For One Worker 18

B.3.4 Explicit Query Bounds And Large-Dimension Optimality 19

C Discretized Setting: Proofs For Multiple-Worker Identification **21**

C.1 Deterministic Identification: Proof for Theorem 4.1 21

C.2 Randomized Identification: Proof for Theorem 4.2 21

C.2.1 Exact Query Characterization On A Complete Network 21

C.2.2 Coverage And Confidence Lower Bounds For Multiple Workers 23

C.2.3 Total Communication With Worker-To-Worker Links 24

D Probabilistic Setting: Formal Setup, Communication Conventions, and Notation **25**

E Probabilistic Setting: Results For Single Worker **27**

E.1 Deterministic Verification 27

E.2 Randomized Verification: Proof for Theorem 5.1 28

F Probabilistic Setting: Results For Multiple Workers **30**

F.1 Deterministic Identification 30

F.2 Randomized Identification: Proof for Theorem 5.2 31

A Discretized Setting: Communication Conventions And Notation

Recall the matrix family is $\mathcal{X} = \mathcal{F}_M^{N_{\text{out}} \times N_{\text{in}}}$ and has size $H = M^{N_{\text{out}} N_{\text{in}}} = 2^N$. In this appendix, a row and a row difference are row vectors. Zero-column queries are allowed: their products vanish, so no query can distinguish unequal matrices. Write $\Delta(\mathcal{S})$ for probability laws on a finite set $\mathcal{S}$, $\mathbf{1}\{E\}$ for the indicator of an event E, and $\mathbb{E}$ for expectation.

Communication conventions. A unique communication token starts at Alice. Only its holder performs a protocol step, and a nonempty self-delimiting frame transfers the token. A fixed traversal may determine the next recipient from the preceding transcript, but any input-dependent recipient choice must be encoded and charged. From either side of a worker cut, its local input, its private tape, and the charged crossing transcript determine crossing endpoints, frame boundaries, and token state. There is no uncharged information in timing, silence, or stopping. These conditions are sufficient for the two-party transcript factorizations used below. Deterministic protocols halt on every input; randomized protocols halt almost surely on every input and never exceed their hard query or transmitted-bit caps.

Alice's A is her private input. Workers obtain all matrix-dependent information through counted local oracle evaluations. Any universal query family or hash table in a protocol description is fixed independently of all matrix inputs. Its selected private index is transmitted when the other party needs it. In contrast, reference-dependent query-game optimizers are computed by Alice, and their actual query vectors are sent to the worker. Assume rational linear-program solutions can be sampled exactly using an almost-surely terminating private-coin procedure. Only local computation uses such a procedure. The encoded upper bounds below have fixed finite coin lengths.

Notations. For reference, a query has a canonical rank code of length

$$b_q = \lceil N_{\text{in}} \log_2 M \rceil$$

in $\mathcal{F}_M^{N_{\text{in}}}$, and a response has a fixed code of length

$$b_{\text{ans}} = N_{\text{out}} \left\lceil \log_2(1 + N_{\text{in}}(M-1)^2) \right\rceil .$$

The latter concatenates fixed-length codes for coordinates in $[0, N_{\text{in}}(M-1)^2]$. A literal addressed exchange can be framed in at most

$$R_t = b_q + b_{\text{ans}} + 2\lceil \log_2(t+1) \rceil + 8$$

bits, including endpoint codes and control flags. A query-optimal finite-game protocol of cap c therefore has a finite implementation with total communication bits $C_{\text{tot}} \le cR_t$. Encoded protocols use fixed routes, for which start/return control costs $O(1)$ per visited worker.

B Discretized Setting: Proofs For Single-Worker Verification

B.1 Deterministic Verification: Proof for Theorem 3.1

The following subsections give the lemmas for proving the theorem. Lemma B.1 proves the matching worst-reference query order, and Algorithm 3 simultaneously attains that order and $O(N)$ bits. For a deterministic communication lower bound, grant both parties their entire matrices for free, which can only help. The transcript-rectangle lower bound in Lemma B.2 with zero error requires $\Omega(N)$ bits on a family of size $H = 2^N$. Thus additional local probes cannot bypass the bit lower bound.

B.2 Separating Queries Characterization

Let $\mathbf{1}$ denote the all-ones row of length N_{in} when used as a vector (the indicator notation has event braces). For a reference A, define $D(A)$ to be the minimum number of columns in a query matrix S such that

$$\{e \in \mathcal{F}_M^{N_{\text{in}}} - a : eS = 0\} = \{0\} \quad \text{for every row } a \text{ of } A. \tag{19}$$

Equivalently, $BS = AS$ implies $B = A$ for every $B \in \mathcal{X}$. Put $D_{\max} = \max_{A \in \mathcal{X}} D(A)$.

Lemma B.1 (Separating families)**.** *The fixed-reference deterministic query optimum is exactly $D(A)$. With $a_0 = \lfloor (M-1)/2 \rfloor$,*

$$\left\lceil \frac{N_{\text{in}} \log(a_0+1)}{\log(1+N_{\text{in}} a_0 (M-1))} \right\rceil \le D_{\max} \le U = \min\left\{ \left\lceil \frac{2N_{\text{in}}}{3} \right\rceil, \left\lceil \frac{16 N_{\text{in}} \log(2M)}{\log(N_{\text{in}}+1)} \right\rceil \right\}, \tag{20}$$

where the upper bound uses one fixed query family reconstructing every matrix in $\mathcal{X}$. Consequently $D_{\max} = \Theta(N_{\text{in}} \log M / \log(N_{\text{in}} M))$.

Proof. Run a deterministic verifier on $B = A$. If the resulting query list does not satisfy (19), a legal nonzero replacement in one row gives exactly the same replies. It follows the same adaptive path and is accepted incorrectly. Conversely, comparing the replies to any separating list verifies A exactly. This proves the fixed-reference assertion.

For the lower bound take every reference row equal to $a_0\mathbf{1}$. A separating schedule must be injective on $\{0, \dots, a_0\}^{N_{\text{in}}}$. Indeed, a collision $xS = yS$ gives a nonzero legal difference $x - y$ because $a_0\mathbf{1} + x - y \in \mathcal{F}_M^{N_{\text{in}}}$. Each measured sum lies in $[0, N_{\text{in}} a_0 (M-1)]$. Counting its possible answer vectors gives

$$(1 + N_{\text{in}} a (M-1))^m \ge (a+1)^{N_{\text{in}}}.$$

This proves the lower bound in (20). Since $a_0 + 1 \ge \sqrt{M}$ and $1 + N_{\text{in}} a_0 (M-1) \le N_{\text{in}} M^2$, it is at least $N_{\text{in}} \log M / [4 \log(N_{\text{in}} M)]$. One hard row suffices, and it can be repeated for any N_{out}.

For the first upper bound on $D_{\max}$, on each coordinate triple use the two queries $(M-1, 1, 0)$ and $(0, M-1, 1)$. A colliding row difference e has $e_2 = -(M-1)e_1$ and $e_3 = (M-1)^2 e_1$. Since e is integral and $|e_j| \le M-1$, this forces $e = 0$. Disjoint triples and basis probes on leftover coordinates use exactly $\lceil 2N_{\text{in}}/3 \rceil$ queries.

For the second upper bound on $D_{\max}$, consider a nonzero integer error with k nonzero coordinates and a uniform binary query. After flipping the random bits associated with negative coefficients, a zero response corresponds to choosing a subset whose positive weights sum to a fixed value. More precisely, W.L.O.G., let the nonzero coefficients be $a_1, \dots, a_k$. Let the query bits q_i be independent and uniform in $\{0, 1\}$. Define

$$r_i = \begin{cases} q_i, & a_i > 0, \\ 1 - q_i, & a_i < 0. \end{cases}$$

Then

$$\sum_{i=1}^{k} a_i q_i = \sum_{i=1}^{k} |a_i| r_i - \sum_{i: a_i < 0} |a_i|.$$

Thus a zero response means that the subset $S = \{i : r_i = 1\}$ satisfies

$$\sum_{i \in S} |a_i| = \sum_{i: a_i < 0} |a_i|.$$

Such subsets S form an antichain: adding an element strictly increases the sum, so two nested subsets cannot have the same sum. An antichain of subsets of k elements has at most $\binom{k}{\lfloor k/2 \rfloor}$ members. Since a binary query selects each of the 2^k subsets equally likely, the probability of getting zero is at most

$$p_k = 2^{-k} \binom{k}{\lfloor k/2 \rfloor} \le \min\{1/2, (k+1)^{-1/2}\}.$$

For the second inequality, $p_{2\ell}/p_{2\ell-2} = (2\ell-1)/(2\ell)$ and $(2\ell-1)(2\ell+1) \le 4\ell^2$ give the even case by induction; $p_{2\ell+1} = [(2\ell+1)/(2\ell+2)] p_{2\ell}$ gives the odd case.

Set $m_0 = \lceil 16 N_{\text{in}} \ln(2M) / \ln(N_{\text{in}}+1) \rceil$. If $m_0 \ge N_{\text{in}}$, the basis suffices. Otherwise sample m_0 independent binary queries. The inequality $m_0 < N_{\text{in}}$ implies $\ln(N_{\text{in}}+1) > 16 \ln(2M)$ and $N_{\text{in}} > 2^{16}$. The nonzero row differences of support at most $\sqrt{N_{\text{in}}}$ number at most $N_{\text{in}} (2M N_{\text{in}})^{\sqrt{N_{\text{in}}}}$, whose logarithm is at most $4\sqrt{N_{\text{in}}} \ln N_{\text{in}}$. Since $\ln N_{\text{in}} \le N_{\text{in}}^{1/4}$ for $N_{\text{in}} \ge 2^{16}$,

$$\frac{m_0 \ln 2}{2} \ge 4 \ln 6 \ln 2 \frac{N_{\text{in}}}{\ln N_{\text{in}}} > 4 \frac{N_{\text{in}}}{\ln N_{\text{in}}} \ge 4\sqrt{N_{\text{in}}} \ln N_{\text{in}}.$$

Their combined failure probability is at most $2^{-m_0/2}$. Every larger-support error has zero probability at most $N_{\text{in}}^{-1/4}$ per query. There are fewer than $(2M)^{N_{\text{in}}}$ such errors, so their combined failure probability is at most

$$(2M)^{N_{\text{in}}} N_{\text{in}}^{-m_0/4} \le (2M)^{-N_{\text{in}}},$$

using $\ln N_{\text{in}} / \ln(N_{\text{in}} + 1) \ge 1/2$. The two failure probabilities sum to less than one. Hence a fixed successful family exists, and exhaustive search can select it independently of all inputs. It separates every legal row difference, simultaneously for all rows. Combining the two universal constructions gives U. Finally, $\min\{N_{\text{in}}, N_{\text{in}} \log M / \log(N_{\text{in}} + 1)\} = \Theta(N_{\text{in}} \log M / \log(N_{\text{in}} M))$, by splitting according to the relative sizes of $\log M$ and $\log N_{\text{in}}$. The ceilings are harmless because the latter ratio is at least one. □

Algorithm 3 Single-worker deterministic protocol (Theorem 3.1)

Input: Alice's private A; worker oracle B; a fixed universal family of U legal queries.
Output: Alice accepts exactly when $B = A$.

1: Alice sends a fixed start frame, transferring the token to the worker.
2: The worker makes the U local probes and recovers the unique matrix in $\mathcal{X}$ consistent with its responses.
3: It sends the matrix's canonical $\lceil N \rceil$-bit index and returns the token.
4: Alice compares that index with the index of A and outputs the equality decision.

Queries: $Q = U = O(N_{\text{in}} \log M / \log(N_{\text{in}} M))$.
Total bits: $C_{\text{tot}} \le \lceil N \rceil + O(1)$.

B.2.1 Communication Lower Bounds For Deterministic Potocols For A Single Worker

Here we prove a lemma that can be used for proving the communication lower bounds for both the deterministic and randomized protocols.

To use the lemma for the deterministic protocols in Theorem 3.1, simply use $\eta = 0$ and $J = 2^N$ to show that the total bits $b = \Omega(N)$.

Lemma B.2 (Equality transcripts and private packing)**.** *Consider a two-party protocol on a finite input family of size J, with a hard cap of b transmitted bits on all inputs. Suppose its mean acceptance probability on uniformly distributed equal inputs is at least $a_0 > 0$, and its acceptance probability on every unequal input pair is at most η. Then*

$$\eta + 1/J \ge a_0^2 2^{-(b+1)}. \tag{21}$$

This inequality remains valid if the coins are exposed to both parties for the analysis. If the coins are private, the mean diagonal acceptance is at least $3/16$, and $\eta \le 1/64$, then

$$J/14 \le (17K)^K, \qquad K \le 2^{b+1}. \tag{22}$$

In particular, the latter conditions imply $b = \Omega(\log_2(\log_2 J + 1))$ for growing J.

Proof. For the first assertion, fix the exposed coins and append a decision bit if necessary. Each accepting transcript determines a rectangle $X_\tau \times Y_\tau$ of input pairs: for a fixed preceding transcript, each message imposes a condition only on its sender's input. There are at most 2^{b+1} accepting transcripts. Their rectangles partition the accepted pairs. Write $D_\tau = X_\tau \cap Y_\tau$ and $a = \sum_\tau |D_\tau|/J$ for the acceptance fraction on equal inputs for these fixed coins. The acceptance probability on independent uniform inputs is at least

$$\sum_\tau \frac{|X_\tau||Y_\tau|}{J^2} \ge \sum_\tau \left(\frac{|D_\tau|}{J}\right)^2 \ge a^2 2^{-(b+1)}.$$

Averaging the coins and applying Jensen's inequality gives a lower bound of $a_0^2 2^{-(b+1)}$. Independent uniform inputs agree with probability $1/J$ and otherwise are accepted with probability at most η, proving (21).

For private coins, acceptance probabilities have the factorization

$$F(x,y)=\sum_{\tau=1}^{K}\alpha_\tau(x)\beta_\tau(y),\qquad 0\le\alpha_\tau(x),\beta_\tau(y)\le 1,\qquad K\le 2^{b+1}.$$

Here the factors are the probabilities that each party's private tape is consistent with its own input and the fixed accepting transcript, including the appended decision bit. Independent tapes give the product. Since the mean of $F(x,x)$ is at least $3/16$, at least $J/14$ inputs satisfy $F(x,x)\ge 1/8$: if their fraction is f, the mean is at most $f+(1-f)/8$. For distinct such inputs x,y, their vectors $\alpha(x)$ and $\alpha(y)$ cannot be within $1/(16K)$ in every coordinate. Otherwise

$$|F(x,x)-F(y,x)|\le\sum_{\tau=1}^{K}|\alpha_\tau(x)-\alpha_\tau(y)|\beta_\tau(x)\le 1/16,$$

whereas diagonal and off-diagonal acceptance give a gap at least $1/8-1/64=7/64$. Partition $[0,1]^K$ into boxes of side $1/(16K)$. At most one of the good input vectors lies in each box, and there are at most $(17K)^K$ boxes. This proves (22). Taking logarithms yields

$$\log_2 J-O(1)\le 2^{b+1}(b+O(1)),$$

and hence the stated logarithmic lower bound. □

B.3 Randomized Verification: Proof for Theorem 3.2

The next subsections prove Theorem 3.2. We first construct the encoded protocol and prove its communication optimality. An path representation and a fixed-prior lower bound then establish the matching query order under $N_{\mathrm{in}}\ge 1/\varepsilon$, including against adaptive two-sided verifiers.

B.3.1 Private-Coin Encoding and Resource Upper Bounds

The local probes in this subsection are ordinary-integer matrix products. Binary strings are used only to encode those products for communication. In particular, hashing a response never replaces its entries by their parities.

Lemma B.3 (A finite private hash family)**.** *Let $\mathcal{Y}$ be a finite family of cardinality $J\ge 2$, and let $0<\eta<1/2$. There is an input-independent finite table of functions from $\mathcal{Y}$ to $\{0,1\}^k$, where $k=\lceil\log_2(4/\eta)\rceil$, such that a uniform table index has collision probability at most η for every unequal pair. The table size is a power of two and is $O(\log(2J)/\eta^2)$. Consequently, Alice can privately select and send an index, receive the selected hash of the other party's input, and test equality with one-sided error at most η using $O(\log(\log J+1)+\log(1/\eta))$ total bits, including fixed framing.*

Proof. Draw K independent uniform functions from $\mathcal{Y}$ to $\{0,1\}^k$, allowing repetitions. For a fixed unequal pair the mean collision probability is $2^{-k}\le\eta/4$. For independent random variables $Z_j\in[0,1]$, the exponential moment inequality $\mathbb{E}\exp(\theta(Z_j-\mathbb{E}Z_j))\le\exp(\theta^2/8)$ and exponential Markov give

$$\Pr\left\{K^{-1}\sum_{j=1}^{K}Z_j-K^{-1}\sum_{j=1}^{K}\mathbb{E}Z_j\ge a\right\}\le\exp(-2Ka^2),\qquad a>0.$$

For completeness, the moment inequality follows by convexly bounding the exponential of a variable in $[0,1]$ by its endpoint chord; the logarithm of the resulting centered Bernoulli moment generating function has value and derivative zero at the origin and second derivative at most $1/4$. Apply this bound with $a=3\eta/4$ and take a union bound over fewer than J^2 unequal pairs. A sufficiently large power of two $K=O(\log(2J)/\eta^2)$ makes the failure probability less than one. Fix a successful table as part of the protocol description. Its $\log_2 K$-bit index can be sampled with a fixed number of private fair bits. Alice transmits this index and receives the k-bit hash value. Equal inputs always have equal hashes. The table itself is fixed before either input is chosen and does not provide a shared random index for free. □

Proof for the upper bounds in Theorem 3.2. Put $m = m_\varepsilon = \min\{U, \lceil \log_M(8/\varepsilon) \rceil\}$. If $m = U$, use the fixed universal reconstructing list from the deterministic proof. Otherwise draw m independent uniform queries from $\mathcal{F}_M^{N_{\text{in}}}$. For any fixed unequal A, B, these queries miss with probability at most $M^{-m} \le \varepsilon/8$. The one-query estimate follows by choosing a nonzero coefficient of one row of $B - A$: after the other query coordinates have been fixed, at most one value of the remaining coordinate can make its dot product zero. Independence gives the estimate for m queries.

In this latter case, draw K independent m-query schedules. The exponential estimate in Lemma B.3, with gap $\varepsilon/8$, and a union bound over fewer than $|\mathcal{X}|^2 = 2^{2N}$ ordered unequal matrix pairs show that a power-of-two table of size $K = O((N+1)/\varepsilon^2)$ can have empirical miss probability at most $\varepsilon/4$ for every pair. Fix such a table, and privately choose a uniform schedule index. Its length is $O(\log_2(N+1) + L)$. The universal-list branch needs no schedule index.

Every response coordinate is an integer between zero and $N_{\text{in}}(M-1)^2$. Encode the complete array of m responses injectively in

$$D_{\text{resp}} = N_{\text{out}} m \left\lceil \log_2\bigl(1 + N_{\text{in}}(M-1)^2\bigr) \right\rceil \tag{23}$$

bits, using the same fixed padding convention at both parties. Since $m \le U \le N_{\text{in}}$, $D_{\text{resp}} = O(N(1 + \log_2(N+1)))$, and therefore $\log_2(D_{\text{resp}} + 1) = O(\log_2(N+1))$. Independently choose a hash table index from Lemma B.3 for the family $\{0,1\}^{D_{\text{resp}}}$, with target collision probability $\eta = \varepsilon/4$. Its hash has $k = \lceil \log_2(16/\varepsilon) \rceil$ bits.

Alice sends the needed schedule and hash indices in a start frame. The worker generates the indicated queries locally, makes its m counted probes, and returns the selected hash of its response array. Alice compares it with the hash of her own array, computed from A. Equal matrices always pass. The schedule misses with probability at most $\varepsilon/4$; conditional on any schedule producing unequal arrays, independence makes the hash collision probability at most $\varepsilon/4$. Thus the miss probability is at most $\varepsilon/2 \le \varepsilon$, and the simultaneous hard bounds are

$$Q \le m_\varepsilon, \qquad C_{\text{tot}} = O(\log_2(N+1) + L).$$

All seed indices, response hashes, and fixed start/return frames are included in this communication bound.

For the exact alternative, the worker uses the universal list to recover its matrix and sends its canonical index in $\mathcal{X}$. This has $Q \le U$ and $C_{\text{tot}} \le \lceil N \rceil + O(1)$, with zero error. Select this alternative when $N \le \log_2(N+1) + L$, and otherwise use hashing. This choice attains $O(\min\{N, \log_2(N+1) + L\})$ bits and still satisfies $Q \le m_\varepsilon$. To check the latter point, for $N \ge 1$, $\log_2(N+1) \le N/3 + 3$. The branch condition implies $L + 3 \ge 2N/3$, and hence

$$\lceil \log_M(8/\varepsilon) \rceil = \left\lceil \frac{L+3}{\log_2 M} \right\rceil \ge \left\lceil \frac{2N_{\text{out}}N_{\text{in}}}{3} \right\rceil \ge U.$$

Thus the exact branch is selected only when $m_\varepsilon = U$. This proves the simultaneous upper bounds used by Algorithm 1. □

B.3.2 Communication Lower Bounds For A Single Worker

Allowing the worker to know its entire matrix can only strengthen a protocol, so a lower bound for equality on $\mathcal{X}$ also applies to the oracle model. Here $J = |\mathcal{X}| = 2^N$. For error at most $\varepsilon \le 1/3$, diagonal acceptance is at least $2/3$. Applying Lemma B.2, Equation (21) gives

$$\varepsilon + 2^{-N} \ge c\, 2^{-b}$$

for an absolute constant $c > 0$. Consequently $b = \Omega(\min\{N, L\})$ whenever this minimum grows. The bounded regime follows because with no communication Alice's output, for a fixed reference, cannot distinguish equality from an unequal worker input with error below $1/2$.

For the private-coin bound, repeat the protocol an absolute constant number of times with independent tapes and take a majority decision. This reduces error from at most $1/3$ to at most $1/64$, with a constant-factor bit increase. Equation (22) in Lemma B.2 then gives the additional lower bound $b = \Omega(\log_2(N+1))$. Together these bounds imply

$$b = \Omega(\min\{N, \log_2(N+1) + L\}),$$

because the quantity inside this minimum is at most $\log_2(N+1) + \min\{N, L\}$. The encoded upper constructions give the matching order. Exact matrix transmission after universal reconstruction gives the deterministic upper bound $O(N)$.

B.3.3 Exact Randomized Query Complexity For One Worker

For the execution path on the uncorrupted input $B = A$ (called honest-path), we establish an honest-path representation that will be used to prove the query lower bound in Theorem 3.2.

We use the following finite game to formalize the honest-path lower bound. For an integer $m \geq 0$, a schedule $S \in \mathcal{F}_M^{N_{\text{in}} \times m}$ contains m query columns. For fixed A, set

$$\beta_m(A) = \min_{\pi \in \Delta(\mathcal{F}_M^{N_{\text{in}} \times m})} \max_{B \in \mathcal{X} \setminus \{A\}} \mathop{\mathbb{E}}_{S \sim \pi} \mathbf{1}\{(B - A)S = 0\}, \qquad \beta_m^\star = \max_{A \in \mathcal{X}} \beta_m(A). \tag{24}$$

The game chooses a distribution of tests against the most difficult fixed corruption. It is a finite linear program; its distribution may depend on Alice's reference. All probabilities below concern private random coins drawn independently of the fixed input matrices. A hard query cap bounds the number of probes on every execution. The finite probability distributions used below can be sampled exactly with almost-surely terminating local computation: their linear programs have rational coefficients and rational optimal solutions. The number of local coin flips is not charged as communication. Thus exact sampling does not require an optimal probability to be dyadic. No efficient algorithm for solving these finite programs is claimed.

Lemma B.4 (Honest-path characterization)**.** *For a fixed reference $A \in \mathcal{X}$ and an integer $m \geq 0$, the optimal one-sided miss probability with at most m probes is $\beta_m(A)$. The optimal maximum of the false-positive and false-negative probabilities is*

$$e_m(A) = \frac{\beta_m(A)}{1 + \beta_m(A)}. \tag{25}$$

These statements allow adaptive probes. In particular, $\beta_0(A) = 1$, $e_0(A) = 1/2$, and zero error is possible precisely when $m \geq D(A)$.

Proof. First consider a one-sided verifier. Couple executions by fixing all its random tapes, including any tapes used by the worker's faithful controller. Follow the execution on the honest matrix $B = A$. List its at most m queries, padding with zero queries if necessary, as the columns of a matrix $S \in \mathcal{F}_M^{N_{\text{in}} \times m}$. If an alternative B satisfies $(B - A)S = 0$, induction over the queries shows that it produces exactly the honest responses. The same tapes therefore lead to the same transcript and acceptance. Consequently its miss probability is at least $\mathbb{P}\{(B - A)S = 0\}$, and maximizing over $B \neq A$ gives a lower bound of $\beta_m(A)$ after minimizing over the induced law of S. Conversely, sample a minimizing law in the definition of $\beta_m(A)$, execute that list, and reject at the first response different from Aq. This is one-sided and has miss probability exactly $\mathbb{P}\{(B-A)S = 0\}$ on alternative B. Finiteness of the query and matrix families ensures that the minimum is attained.

For a general two-sided verifier, rejecting at a first discrepancy cannot increase either error. Let p be its rejection probability on A. For each tape that rejects along the honest path, replace the entire execution by immediate rejection. This preserves the null error and cannot increase an alternative's error. Conditional on the other tapes, the verifier accepts along its honest path, and those paths induce a law on query lists. If $p < 1$, some alternative is invisible with conditional probability at least $\beta_m(A)$. Hence every two-sided verifier has maximum error at least

$$\max\{p, (1 - p)\beta_m(A)\} \geq \frac{\beta_m(A)}{1 + \beta_m(A)}.$$

The case $p = 1$ already has error one. Equality is attained by rejecting without probing with probability $\beta_m(A)/(1 + \beta_m(A))$, and otherwise running the minimizing one-sided verifier.

With no probes every alternative is invisible, proving the zero-query claims. If $\beta_m(A) = 0$, every alternative is invisible with probability zero under a minimizing law. There are finitely many alternatives, so with probability one the sampled list separates A from all of them. Some such list therefore exists and $m \geq D(A)$. A separating list gives the converse. Finally, almost-sure termination causes no difficulty in fixing tapes: for each fixed A there are only finitely many possible worker matrices, so a finite union of exceptional null sets still has probability zero. □

Passing to the worst reference. The matrix family $\mathcal{X}$ is finite, and $\beta_m(A)$ is nonincreasing in m because zero-query padding is allowed. The supervisor knows A and may choose the optimal law for

that reference. Since $x/(1+x)$ is increasing for $x \geq 0$, the worst-reference error at cap m is exactly

$$\max_{A \in \mathcal{X}} e_m(A) = \frac{\beta_m^\star}{1+\beta_m^\star}.$$

For $0 < \varepsilon \leq 1/3$, the least cap that works for every reference is therefore

$$\min\left\{ m \in \mathbb{Z}_{\geq 0} : \beta_m^\star \leq \frac{\varepsilon}{1-\varepsilon} \right\}. \tag{26}$$

Equivalently, one may first find the least feasible cap for each A and then maximize over A: monotonicity makes the two operations agree. This is a query characterization. A reference-dependent minimizing list is not shared with the worker for free. Sending its actual queries and receiving the actual responses implements the characterization with the literal communication cost in Appendix A; it does not establish optimal encoded communication.

B.3.4 Explicit Query Bounds And Large-Dimension Optimality

Lemma B.5 (A legal-alphabet random probe)**.** *Let $E = B - A \neq 0$ have rank ρ over the real numbers. A uniform $q \in \mathcal{F}_M^{N_{\text{in}}}$ obeys*

$$\mathbb{P}\{Eq = 0\} \leq M^{-\rho}.$$

A uniform nonzero query obeys

$$\mathbb{P}\{Eq = 0\} \leq \alpha_\rho := \frac{M^{N_{\text{in}}-\rho} - 1}{M^{N_{\text{in}}} - 1}. \tag{27}$$

For $N_{\text{in}} \geq 2$, put $\alpha = \alpha_1$. Independent nonzero queries have worst-alternative miss exactly α^m for every reference A.

Proof. Choose ρ independent rows of E and ρ pivot columns making their corresponding square submatrix invertible. Once the other $N_{\text{in}} - \rho$ coordinates of q are fixed, these equations determine the pivot coordinates uniquely over the reals. There is thus at most one legal choice of pivot coordinates for each assignment of the other coordinates. At most $M^{N_{\text{in}}-\rho}$ cube points are in the kernel. The zero vector is one such point; removing it from both the numerator and denominator gives (27).

For every reference A, changing a single entry by $+1$ or -1 is possible within $\mathcal{F}_M$. This gives an error whose kernel is exactly $q_j = 0$ for one coordinate j, and hence attains α_1. Independence raises this miss probability to the mth power. More generally, errors supported on ρ separate diagonal coordinates attain the rank bounds when the dimensions and chosen matrices permit these changes. □

The full-cube probe therefore gives, for any one-sided target miss $0 < \eta < 1$, a cap of

$$\min\left\{D(A), \lceil \log_M(1/\eta) \rceil\right\}. \tag{28}$$

Here and below a cap $D(A)$ uses a separating list; the other branch uses independent random probes. One may replace $D(A)$ by the universal size U from Lemma B.1. For $N_{\text{in}} \geq 2$, nonzero probes improve the random term to $\lceil \log(1/\eta)/\log(1/\alpha) \rceil$. For $N_{\text{in}} = 1$, one nonzero probe is exact. For two-sided error ε, the honest-path lemma permits $\eta = \varepsilon/(1-\varepsilon)$ in these query-only bounds. The miss probability of an IID construction is not, by itself, a lower bound against other distributions or adaptive protocols.

Lemma B.6 (Fixed-prior lower bounds)**.** *Let $a = \lfloor (M-1)/2 \rfloor$ and choose A with one row equal to $a\mathbf{1}$; the remaining rows may be equal to this row. Define*

$$T = (a+1)^{N_{\text{in}}}, \qquad K_0 = 1 + N_{\text{in}} a(M-1), \qquad [x]_+ = \max\{x, 0\}.$$

For $N_{\text{in}} \geq 2$, set $K = \min\{M^m, N_{\text{in}}\}$ and write $N_{\text{in}} = uK + v$ with integers $u \geq 1$ and $0 \leq v < K$. Then

$$\beta_m(A) \geq \frac{v(u+1)u + (K-v)u(u-1)}{N_{\text{in}}(N_{\text{in}}-1)} \geq \left[\frac{N_{\text{in}} M^{-m} - 1}{N_{\text{in}} - 1} \right]_+, \tag{29}$$

$$\beta_m(A) \geq \left[\frac{TK_0^{-m} - 1}{T - 1}\right]_+. \tag{30}$$

The second bound also holds for $N_{\text{in}} = 1$. Consequently, a one-sided target miss η requires

$$m \geq \log_M \left(\frac{N_{\text{in}}}{1 + \eta(N_{\text{in}} - 1)}\right) \qquad (N_{\text{in}} \geq 2), \tag{31}$$

$$m \geq \frac{\log T - \log(1 + \eta(T - 1))}{\log K_0}. \tag{32}$$

For two-sided error ε, substitute $\eta = \varepsilon/(1 - \varepsilon)$.

Proof. The key is to choose a distribution of legal alternatives before the verifier draws any coins. For the first distribution, choose an unordered pair $\{j, k\}$ uniformly and add $e_j - e_k$ to the designated reference row, where e_j is the jth standard basis row vector. These are legal nonzero errors because $a \geq 1$ and $a + 1 \leq M - 1$. For any fixed honest-path query matrix S, coordinate j has a length-m codeword given by row j of S, with at most M^m possible codewords. The pair error is invisible exactly when its two codewords agree.

If the nonempty codeword cells have sizes $z_1, \ldots, z_K$, padding with empty cells as needed, the collision fraction is $\sum_{\ell=1}^{K} z_\ell(z_\ell - 1)/[N_{\text{in}}(N_{\text{in}} - 1)]$. Moving one coordinate from a cell to a cell smaller by at least two strictly decreases the numerator. Its minimum is therefore attained by v cells of size $u + 1$ and $K - v$ cells of size u, proving the first expression in (29). Alternatively, $\sum_\ell z_\ell^2 \geq N_{\text{in}}^2/K$ gives the displayed simpler bound. Average over the verifier's tapes only after this pointwise bound on each query list. Since a maximum over alternatives dominates their fixed-prior average, the bound applies to $\beta_m(A)$, including adaptive verifiers by Lemma B.4.

For the second distribution, draw independent uniform vectors $X, Y \in \{0, \ldots, a\}^{N_{\text{in}}}$ conditioned on $X \neq Y$, and add $X - Y$ to the designated row. Its entries lie between zero and $2a \leq M - 1$, so this also defines a fixed distribution of legal nonzero alternatives. For a fixed S, each coordinate of XS is an integer in $[0, N_{\text{in}}a(M - 1)]$. Thus XS has at most K_0^m possible values. If their probabilities are p_w, then

$$\mathbb{P}\{XS = YS\} = \sum_w p_w^2 \geq K_0^{-m}.$$

The event $X = Y$ has probability $1/T$ and always gives a collision. Conditioning it away yields

$$\mathbb{P}\{(X - Y)S = 0 \mid X \neq Y\} \geq \frac{K_0^{-m} - 1/T}{1 - 1/T}.$$

Truncation at zero and averaging over tapes proves (30). Rearrangement gives the two inverse bounds; Lemma B.4 gives their two-sided versions. □

Query lower bounds in Theorem 3.2. Let $\eta = \varepsilon/(1 - \varepsilon)$. By the honest-path characterization and the fixed-prior bound, every two-sided verifier valid for all references requires

$$m \geq \log_M \frac{1}{\eta + (1 - \eta)/N_{\text{in}}}.$$

Since $0 < \varepsilon \leq 1/3$, we have $0 < \eta < 1$. The condition $N_{\text{in}} \geq 1/\varepsilon$ therefore gives

$$\eta + \frac{1 - \eta}{N_{\text{in}}} \leq \eta + (1 - \eta)\varepsilon = 2\varepsilon.$$

Thus $m \geq \lceil \log_M(1/(2\varepsilon)) \rceil$. Conversely, $\lceil \log_M(1/\varepsilon) \rceil$ independent uniform probes have one-sided miss at most ε for every unequal pair. This proves the two displayed query bounds in the theorem. Their integer endpoints differ by at most one because $\log_M 2 < 1$.

For completeness, put $x = \log_M(1/\varepsilon) > 0$. When $x \geq 2$, the lower bound is at least $x - 1 \geq x/2$ and hence at least $\lceil x \rceil/3$. When $x < 2$, a zero-query verifier cannot achieve error below $1/2$, so its cap is at least one and hence at least $\lceil x \rceil/2$. These observations give a uniform $\Omega(\lceil x \rceil)$ bound, including when M grows. Algorithm 1 has $m_\varepsilon \leq 3\lceil x \rceil$ and the optimal communication order proved above. It therefore attains both orders simultaneously. □

C Discretized Setting: Proofs For Multiple-Worker Identification

C.1 Deterministic Identification: Proof for Theorem 4.1

Fix a reference A. On the all-honest execution of a deterministic protocol, let g be its output and let S_i be all queries made at worker i. For every $i \neq g$, the list S_i must separate A from every allowed alternative. Otherwise place a colliding matrix at worker i; all local replies and, by faithfulness, all messages and subsequent queries remain the same, so the protocol returns the wrong index g. Each of the $t-1$ nonfallback workers therefore requires at least $D(A)$ probes. This argument permits arbitrary peer messages because controllers have no matrix information beyond their probe responses.

Conversely, a separating list at each of $t-1$ workers identifies the first discrepant worker, or permits the last worker to be inferred. Thus the exact fixed-reference query cap is $(t-1)D(A)$, and the worst-reference cap is $(t-1)D_{\max}$. Lemma B.1 gives the claimed order. To attain it simultaneously with optimal communication, use the universal reconstructing family and Algorithm 4. The network communication proof in Appendix C.2.3 gives its matching $\Theta(tN)$ bit cost, completing Theorem 4.1.

Algorithm 4 Deterministic complete-network identification (Theorem 4.1)

Input: Alice's A; $t \geq 2$ faithful worker oracles; the universal U-query family.
Output: The index of the unique corrupted worker on every valid tuple.

1: **for** $i = 1, \ldots, t-1$ in a fixed traversal **do**
2: Run Algorithm 3 with worker i over its Alice–worker link.
3: If its matrix index differs from Alice's index, return i.
4: **end for**
5: **return** t.

Queries: $Q \leq (t-1)U$.
Total bits: $C_{\text{tot}} = O(tN)$, including start/return frames on all used links.
The complete network permits peer edges; this construction uses only the fixed Alice–worker traversal.

C.2 Randomized Identification: Proof for Theorem 4.2

We first reduce adaptive queries to honest-path plans and derive fixed-prior lower bounds. We then prove the matching query rate and the private communication order in Theorem 4.2.

More precisely, the query proof in Section C.2.2 establishes (11) for every $0 < \varepsilon \leq 1/3$ with $N_{\text{in}} \geq 1/\varepsilon$, and the communication arguments in Section C.2.3 establish (12). Algorithm 2 attains both orders simultaneously. The optimality claims are asymptotic with absolute constants throughout the stated dimension regime.

C.2.1 Exact Query Characterization On A Complete Network

We give the plan representation used in the query lower bound of Theorem 4.2. Recall that communication may use every edge of the complete network, but every worker's access to its matrix remains limited to its counted local probes. The reference A is known to the supervisor. Exactly one of the $t \geq 2$ matrices differs from A, and the objective is its index.

For the supporting randomized query characterization, a *null plan* of integer cap $c \geq 0$ is an ordered list of at most c addressed queries (i, q), together with a fallback index $g \in [t]$. Let $\mathcal{P}_c$ be the finite set of these plans. For a plan $p \in \mathcal{P}_c$, write $S_i(p)$ for the matrix of query columns addressed to worker i and $g(p)$ for its fallback. Define

$$\mathcal{E}_c(A) = \min_{\pi \in \Delta(\mathcal{P}_c)} \max_{\substack{i \in [t] \\ B \in \mathcal{X} \setminus \{A\}}} \mathop{\mathbb{E}}_{p \sim \pi} \left[\mathbf{1}\{i \neq g(p)\} \mathbf{1}\{(B-A) S_i(p) = 0\} \right], \qquad \mathcal{E}_c^\star = \max_A \mathcal{E}_c(A). \tag{33}$$

The fallback and all query lists may be correlated. A plan is executed until a discrepancy identifies a worker, or its fallback is returned.

Lemma C.1 (Oracle-only centralized simulation)**.** *Every complete-network protocol has a centralized simulation with the same distribution of outputs and the same number of addressed probes. For query lower bounds, this simulation may ignore the communication cost of its internal calculations.*

Proof. A centralized simulator starts with A and samples all parties' independent private tapes. It stores their initial controller states and simulates every message and local computation. By the oracle-only assumption, controller states contain no matrix-dependent information before a probe, and subsequent dependence on a matrix comes only from responses already obtained. Whenever a controller requests a new local query q at worker i, the simulator makes that same addressed query and supplies its actual answer to the simulated controller. Induction over protocol steps shows that states, messages, oracle requests, and outputs agree in distribution with the original protocol. This argument neither grants free access to a worker matrix nor asserts a communication-preserving reduction. It only shows that peer messages cannot provide information beyond the collected oracle answers for purposes of a query lower bound. □

Lemma C.2 (Finite-plan minimax identity)**.** *For fixed A and total addressed-probe cap c, the exact optimal identification error on a complete network is $\mathcal{E}_c(A)$. In the finite game, a plan consists of at most c ordered addressed queries and a fallback index $g \in \{1, \dots, t\}$. Its sublist at worker i is S_i, which may be empty. The error indicator for corruption $B \neq A$ at index i is*

$$\mathbf{1}\{i \neq g\}\mathbf{1}\{(B - A)S_i = 0\}.$$

Zero error is possible precisely when $c \geq (t-1)D(A)$.

Proof. Use Lemma C.1 and modify the centralized verifier to return i at its first discrepancy from Aq at worker i. Under the exactly-one promise this answer is correct, so the modification cannot increase error or queries. Fix all tapes and follow the synthetic all-honest execution. Termination and the hard resource cap are required on all tuples, including this invalid all-honest tuple; its terminal index is a fallback g. Its ordered addressed queries form a plan of length at most c.

For a fixed alternative at worker i, if $(B - A)S_i = 0$, every answer along this path agrees with the all-honest answer. The execution reaches the same fallback and is wrong exactly when $i \neq g$. Otherwise it encounters a first discrepancy at worker i and identifies that worker correctly. Thus the modified verifier's error is exactly the expectation of the displayed indicator under its induced distribution on plans. The original verifier has no smaller error, proving the finite-game lower bound. Exceptional nonterminating tapes have probability zero; there are finitely many input tuples, so they can be discarded simultaneously.

Conversely, Alice samples any law on plans using her private coins, sends each actual query to its designated worker, and compares the response with Aq. She returns the first discrepant index and otherwise returns g. These Alice–worker edges exist in the complete network. The error is exactly the game's indicator, proving attainability of its minimum. This implementation has a hard cap of c queries on every tuple, including tuples outside the correctness promise. Literal transmission supplies the separately stated communication upper bound; no reference-dependent plan is shared for free.

At zero error, finiteness of the valid input family implies that, with probability one under a minimizing law, a plan is correct on every valid input. For each $i \neq g$, its list S_i must therefore separate A from every $B \neq A$, requiring at least $D(A)$ probes. Summing over the $t - 1$ such indices gives $c \geq (t-1)D(A)$. Conversely, use a separating list at each of $t - 1$ workers and infer the last worker if all those tests pass. □

Worst-reference cap and feasible allocations. As in the single-worker game, finiteness of $\mathcal{X}$, knowledge of A at the supervisor, and monotonicity in c give the exact cap

$$\min\{c \in \mathbb{Z}_{\geq 0} : \mathcal{E}_c^\star \leq \varepsilon\}. \tag{34}$$

The maximizing reference is chosen before any coins are sampled. The law of the ordered plan and its fallback may correlate queries across workers; replacing this law by independent single-worker tests need not preserve its optimum.

A useful feasible law assigns $t-1$ testing slots and one fallback slot to a uniformly random permutation of the worker indices. If slot j uses a minimizing single-worker list of length m_j, with $\sum_{j=1}^{t-1} m_j \leq c$, then

$$\mathcal{E}_c(A) \leq \frac{1}{t} \sum_{j=1}^{t-1} \beta_{m_j}(A). \tag{35}$$

Indeed, a fixed corrupted worker occupies each slot with probability $1/t$; the fallback slot never causes an error, and a testing slot has miss at most its $\beta_{m_j}(A)$. Zero-length testing slots are allowed, with $\beta_0(A) = 1$. A simpler fixed-order construction tests $t-1$ workers one-sidedly and infers the last. It yields a total query cap

$$(t-1)\min\left\{D(A), \lceil \log_M(1/\varepsilon) \rceil\right\}. \tag{36}$$

There is no union-bound penalty over the honest workers: their exact responses always match.

C.2.2 Coverage And Confidence Lower Bounds For Multiple Workers

Lemma C.3 (Coverage and fixed-prior bounds)**.** *Every protocol identifying the corrupted worker with error at most ε and a hard total cap c satisfies*

$$c \geq t(1-\varepsilon) - 1. \tag{37}$$

For the interior reference and T, K_0 of Lemma B.6, its optimal error additionally obeys

$$\mathcal{E}_c(A) \geq \frac{t-1}{t(N_{\text{in}}-1)}[N_{\text{in}} M^{-c/(t-1)} - 1]_+ \qquad (N_{\text{in}} \geq 2), \tag{38}$$

$$\mathcal{E}_c(A) \geq \frac{t-1}{t(T-1)}[T K_0^{-c/(t-1)} - 1]_+. \tag{39}$$

In particular, an error target ε requires

$$c \geq (t-1)\log_M \frac{N_{\text{in}}}{1 + t\varepsilon(N_{\text{in}}-1)/(t-1)} \qquad (N_{\text{in}} \geq 2), \tag{40}$$

$$c \geq \frac{t-1}{\log K_0}\left[\log T - \log\big(1 + t\varepsilon(T-1)/(t-1)\big)\right]. \tag{41}$$

Proof. For coverage, fix A and any one legal matrix $B \neq A$, and place B at a uniformly random worker index independently of the verifier's coins. For each fixed null plan, at most c different workers receive a query. A corruption at an unqueried worker leaves the entire path unchanged, and at most one such index equals its fallback. Even granting success at every queried worker, the average success is at most $(c+1)/t$. Average over tapes and compare with success at least $1-\varepsilon$ on every valid input to obtain (37).

For the confidence bounds, place a corruption drawn from either fixed prior in Lemma B.6 at a uniformly random worker, independently of all tapes. On a deterministic null plan with fallback g, let m_i be its number of queries to worker i. The pair prior gives average error at least

$$\frac{1}{t}\sum_{i \neq g}\left[\frac{N_{\text{in}} M^{-m_i} - 1}{N_{\text{in}} - 1}\right]_+.$$

The summand is convex and nonincreasing as a function of the real variable $m_i \geq 0$: it is the maximum of zero and a positive multiple of an exponential minus a constant. Since $\sum_{i \neq g} m_i \leq c$, Jensen's inequality gives (38). For the cube prior, the same argument uses the convex nonincreasing function $[T K_0^{-m_i} - 1]_+/(T-1)$ and proves (39). Both priors were chosen before the verifier's coins and worker allocations, so averaging over plans preserves the bounds. Rearrangement yields the inverse inequalities. □

Query bounds in Theorem 4.2. Set $\gamma = t\varepsilon/(t-1)$. Then $\gamma \leq 2\varepsilon \leq 2/3$. With $N_{\text{in}} \geq 1/\varepsilon$, the denominator in the pair inverse bound satisfies

$$\gamma + \frac{1-\gamma}{N_{\text{in}}} \leq \gamma + (1-\gamma)\varepsilon \leq 3\varepsilon - 2\varepsilon^2 \leq 3\varepsilon.$$

Consequently, every protocol valid for all references has total cap

$$c \geq (t-1)\log_M \frac{1}{3\varepsilon}.$$

Independently, coverage gives $c \geq t(1-\varepsilon) - 1 \geq (t-1)/3$ for $t \geq 2$ and $\varepsilon \leq 1/3$. Put $x = \log_M(1/\varepsilon)$. If $x \geq 2$, then $\log_M 3 \leq 1$ yields $c \geq (t-1)(x-1) \geq (t-1)\lceil x \rceil/3$. If $x < 2$, then $\lceil x \rceil \leq 2$ and coverage gives $c \geq (t-1)\lceil x \rceil/6$. In all cases,

$$\frac{t-1}{6} \lceil \log_M(1/\varepsilon) \rceil \leq Q_t^{\mathrm{rand}}(\varepsilon) \leq (t-1) \lceil \log_M(1/\varepsilon) \rceil .$$

The upper bound uses independent uniform probes at each of $t-1$ workers as in (36); only the unique corrupted worker can be missed. This establishes the matching query order with absolute constants. These arguments allow arbitrary peer messages through the centralized simulation and arbitrary adaptive allocations through their honest-path plans.

For simultaneous optimal communication, Algorithm 2 instead uses the encoded one-sided local tests. Its query cap is at most $(t-1)m_\varepsilon \leq 3(t-1)\lceil \log_M(1/\varepsilon) \rceil$, still of the optimal order. The following communication proof establishes its matching total-bit bound. □

C.2.3 Total Communication With Worker-To-Worker Links

Communication upper bounds in Theorems 4.1 and 4.2. Use a fixed traversal of workers $1, \ldots, t-1$ over the available Alice–worker edges. For the deterministic construction, each visited worker uses the universal reconstructing list and sends its canonical matrix index to Alice. Stop at the first unequal index, and otherwise output t. This has

$$Q \leq (t-1)U, \qquad C_{\mathrm{tot}} = O(tN).$$

For positive target error, apply Algorithm 1 to each visited worker, stopping at the first rejection and otherwise outputting t. All local tests are one-sided. Hence honest workers always pass, a corruption at the fallback worker is identified exactly, and only a miss at the unique corrupted tested worker can cause an error. A target ε per test suffices; there is no need to replace it by ε/t. Summing the local hard caps gives the simultaneous upper bounds

$$Q \leq (t-1)m_\varepsilon, \qquad C_{\mathrm{tot}} = O(t \min\{N, \log_2(N+1) + L\}) .$$

Each seed copy and every used edge is charged. The fixed traversal needs only $O(t)$ control bits under the stated framing convention; its next recipient is input independent and determined by the already decoded transcript. Both constructions terminate with the same hard caps on invalid tuples as well as on valid ones. □

Communication lower bounds in Theorems 4.1 and 4.2. Let C be an all-input hard cap on total bits for a protocol with exactly-one identification error at most $\varepsilon \leq 1/3$. For lower bounds we may give each controller its matrix explicitly; this relaxes the probe restriction. Write $J = |\mathcal{X}| = 2^N$. The all-input termination and cost convention makes executions on all-equal tuples well-defined and bounded, even though correctness is required only when exactly one worker differs from Alice.

First obtain the confidence lower bound. Select a uniformly random worker I. Across the cut isolating worker I, one side holds a matrix y and the other side holds x, simulating Alice and all the remaining workers on x. Declare equality when the network output differs from I. If $x \neq y$, the network input is valid, so equality is accepted with probability at most ε. If $x = y$, the network execution is independent of I and its output belongs to $[t]$, so mean diagonal acceptance is $1 - 1/t \geq 1/2$.

Expose the original random tapes and I to both sides for this confidence argument only. Each peer bit crosses two isolated-worker cuts and each Alice–worker bit crosses one, so the mean cut traffic, averaged over a uniform diagonal input, tapes, and I, is at most $2C/t$. Truncate the cut simulation before its crossing transcript would exceed $\max\{1, \lceil 16C/t \rceil\}$ bits, declaring inequality on truncation. Markov's inequality loses at most $1/8$ of mean diagonal acceptance and cannot increase off-diagonal acceptance. The charged, decodable cut transcript makes this a genuine two-party protocol with hard cap $O(C/t + 1)$: cut endpoints, frame boundaries, and token state are not extra uncharged information. Equation (21) therefore yields

$$\varepsilon + 2^{-N} \geq c\, 2^{-c'(C/t+1)} \tag{42}$$

for absolute constants $c, c' > 0$.

We also need a bound that remains nonzero for bounded parameters. Fix distinct matrices x, y and an all-equal execution on x. At most $2C$ workers are endpoints of transmitted bits. Replacing the

matrix at an untouched worker by y leaves the execution unchanged: the token and charged framing exclude input-dependent unscheduled sends, silence, timing, or routing as free signals. At most one untouched worker is named correctly by that execution. Averaging a uniform corrupted index and the tapes therefore gives

$$1 - \varepsilon \le \frac{2C+1}{t}.$$

Thus $C = \Omega(t)$ for $t \ge 2$ and $\varepsilon \le 1/3$. Combining this fact with (42) gives $C = \Omega(t \min\{N, L\})$. Setting $\varepsilon = 0$ gives $C = \Omega(tN)$, including deterministic protocols and randomized zero-error protocols with an all-input hard bit cap.

For the additional private-coin lower bound, amplify with a constant number of independent runs and majority output so that the error is at most $1/64$. A majority means an index appearing on more than half the runs; in the absence of one, use a fixed fallback index. Fixed reset/control frames cost at most $O(t)$ per run. By the participation bound just proved, the amplified all-input cost is therefore $O(C)$.

For the amplified protocol, let a_i be the mean probability of outputting index i on a uniform all-equal matrix, and let T_i be the mean number of bits crossing worker i's cut on those inputs. These means average over all private tapes. Since $\sum_{i=1}^{t} a_i = 1$, at least $t-1$ indices satisfy $a_i \le 3/4$. For each such index keep the cut fixed, rather than revealing a random cut to the parties. Its two sides have independent private tapes, diagonal equality acceptance at least $1/4$ on average, and off-diagonal acceptance at most $1/64$.

Truncate that two-party simulation after $\max\{1, \lceil 16T_i \rceil\}$ crossing bits and declare inequality on truncation. Markov's inequality loses at most $1/16$ of mean diagonal acceptance, leaving at least $3/16$; off-diagonal acceptance can only decrease. The private packing part of Lemma B.2 now gives $T_i = \Omega(\log_2(N+1))$ for growing N. Summing over at least $t-1$ such cuts and using $\sum_i T_i \le 2O(C)$ yields $C = \Omega(t \log_2(N+1))$. The participation bound handles bounded N.

Finally combine the private dimension and confidence terms. Since

$$\min\{N, \log_2(N+1) + L\} \le \log_2(N+1) + \min\{N, L\},$$

their maximum is at least half the desired expression. This proves

$$C = \Omega(t \min\{N, \log_2(N+1) + L\}),$$

matching the upper bound while allowing all worker-to-worker edges. The exact lower and upper bounds similarly give $\Theta(tN)$. □

D Probabilistic Setting: Formal Setup, Communication Conventions, and Notation

Matrices, probes, and a separation promise. In this appendix $N_{\text{out}} \ge 2$ is the output space size and $N_{\text{in}} \ge 1$ is the input space size. (The case $N_{\text{out}} = 1$ has only one possible matrix and no corrupted alternative in this setting, and thus is excluded below.) For an integer $k \ge 1$, write

$$\mathcal{S}_k = \{x \in \mathbb{R}^k : x_i \ge 0 \ (i \in [k]),\ \sum_{i=1}^{k} x_i = 1\}.$$

The reference A and every worker matrix belong to $\mathcal{X}_{\text{p}} = (\mathcal{S}_{N_{\text{out}}})^{N_{\text{in}}}$, with each column being a probability distribution. A legal query $q \in \mathcal{S}_{N_{\text{in}}}$ returns the exact distribution Bq. All arithmetic and local comparisons are exact ordinary real operations. An oracle call returns the full distribution to its local faithful controller, while this does not transmit an arbitrary real to Alice for free.

For probability vectors $p, p' \in \mathcal{S}_{N_{\text{out}}}$ define total variation by $\text{TV}(p, p') = \frac{1}{2}\|p - p'\|_1$, and set

$$\delta(A, B) = \max_{j \in [N_{\text{in}}]} \text{TV}(A_{\cdot j}, B_{\cdot j}).$$

A column at distance $\nu^* > 0$ implies $\delta(A, B) \ge \nu^*$. It does not imply that only one column changes: changes in several columns may cancel under a mixture query. The finite communication results require a *known* threshold $0 < \nu \le \nu^*$. The verification promise is $B = A$ or $\delta(A, B) \ge \nu$;

identification has exactly one worker with distance at least ν, with all others equal to A. The threshold is a lower bound, not a promise that the maximum distance equals precisely ν.

Merely assuming $B \neq A$, with no supplied positive separation, is a different model. There is then no uniform finite worst-case bit cap for private-coin verification or identification at error at most $1/3$, regardless of the number of local probes. We prove this obstruction below. Thus a known gap is necessary for the finite bit bounds stated here, and it is not implicit in the existence of an unknown positive ν^*. The matching communication orders are proved for $0 < \nu \leq 1/8$. For every $0 < \nu \leq 1$ the constructions still give the upper bounds specified below, but we do not claim matching communication lower bounds throughout the larger-gap range.

Private coins, communication, and the two resource optima. Alice knows her private reference A. Workers learn about their stored matrices only through counted local probes and may encode the results. Their controllers execute the protocol faithfully even when the matrix is corrupted. All matrices are fixed before the coins are drawn. Every random choice uses private fair-bit tapes that are initially independent. An execution reads finitely many bits and terminates almost surely on each input. No continuously distributed exact random real or free public coins are used. Local exact computation and input-independent preprocessing are unrestricted. Fixed tables are protocol data, but transmitting their chosen random indices costs bits.

We use exactly the charged token and cut-simulation convention of the discretized setting in Appendix A, including endpoints, framing, routing, and control. Only binary messages cross communication links; timing and silence provide no free information. Every protocol obeys its hard query and bit caps on all inputs, including all-equal tuples outside the identification promise. Alice outputs while holding the token and returns some index in $[t]$ on invalid tuples. All worker–worker edges are available for the multiworker results, and every traversed edge is charged. These are total-traffic bounds, rather than bounds only on Alice's incident communication.

Write Q for the total number of local worker-probes and C_{tot} for total transmitted bits. Each common round contains t probes, while the results below optimize flexible, individually addressed probes. For $t = 1$ we mean equality verification, and for $t \geq 2$ we mean exactly-one identification. Let $Q^{\mathrm{det}}_{\mathrm{p},t}(\nu)$ and $C^{\mathrm{det}}_{\mathrm{p},t}(\nu)$ be the optimal worst-case deterministic caps for these tasks. Let $Q^{\mathrm{rand}}_{\mathrm{p},t}(\nu,\varepsilon)$ and $C^{\mathrm{rand}}_{\mathrm{p},t}(\nu,\varepsilon)$ be their private-coin counterparts. The optima range over all references and allowed worker matrices and optimize each resource separately among protocols with finite hard resource caps. An explicitly stated pair of achievable bounds belongs to a single protocol. In verification, both error types are at most ε; all exhibited randomized tests are one-sided and always accept an honest worker. In identification, ε bounds the probability of naming the wrong worker. Unless a zero-error assertion is explicit, $0 < \varepsilon \leq 1/3$ and $L = \log_2(1/\varepsilon)$.

Define

$$D_{\mathrm{p}} = N_{\mathrm{in}}(N_{\mathrm{out}} - 1), \qquad N_\nu = D_{\mathrm{p}} \log_2(1/\nu), \qquad \overline{N}_\nu = D_{\mathrm{p}} \log_2(4e/\nu),$$
$$F_\nu(u) = \log_2(N_\nu + 1) + \log_2(1/u) \quad (0 < u \leq 1/3). \tag{43}$$

Here e is the base of the natural logarithm, and $F_\nu(\varepsilon) = \log_2(N_\nu + 1) + L$. N_ν and F_ν will be used for $0 < \nu \leq 1/8$, where $\overline{N}_\nu = \Theta(N_\nu)$ and $N_\nu \geq 3N_{\mathrm{in}}$. The quantity D_{p} is the number of free real coordinates in a probability matrix; N_ν measures its description size at resolution ν. All logarithms in bit bounds are base two, and implicit constants are absolute. If protocol parameters must have finite descriptions, a supplied rational lower bound on the true gap suffices. Exact local arithmetic at the very fine resolutions below is an access-model assumption, not a numerical-stability claim.

A finite matrix label. The deterministic protocols use the following input-independent quantization, denoted Quant_ν. Set $G_\nu = \lceil 2(N_{\mathrm{out}} - 1)/\nu \rceil$. For each probability column p put

$$z_a = \lfloor G_\nu p_a \rfloor \quad (a < N_{\mathrm{out}}), \qquad z_{N_{\mathrm{out}}} = G_\nu - \sum_{a < N_{\mathrm{out}}} z_a.$$

The nonnegative integers z_a sum to G_ν. Encode this weak composition by its index in a fixed enumeration of the $\binom{G_\nu + N_{\mathrm{out}} - 1}{N_{\mathrm{out}} - 1}$ possibilities. Concatenating the N_{in} fixed-length column indices defines $\mathsf{Quant}_\nu(B)$, of length

$$b_\nu = N_{\mathrm{in}} \left\lceil \log_2 \binom{G_\nu + N_{\mathrm{out}} - 1}{N_{\mathrm{out}} - 1} \right\rceil. \tag{44}$$

The worker first obtains its columns by local basis probes. This compact composition code, rather than separately padded coordinates, is essential to the stated dimension dependence. Its separation property and bit cost are proved next.

E Probabilistic Setting: Results For Single Worker

E.1 Deterministic Verification

Theorem E.1 (Optimal deterministic probability verification)**.** *For $N_{\mathrm{out}} \geq 2$, $N_{\mathrm{in}} \geq 1$, and a supplied threshold $0 < \nu \leq 1/8$, the worst-case optimal resource bounds are*

$$Q^{\mathrm{det}}_{\mathrm{p},1}(\nu) = N_{\mathrm{in}} \qquad \textit{number of queries}, \tag{45}$$

$$C^{\mathrm{det}}_{\mathrm{p},1}(\nu) = \Theta(N_\nu) \qquad \textit{total communication bits}. \tag{46}$$

Algorithm 5 attains both bounds simultaneously, with $Q = N_{\mathrm{in}}$ and $C_{\mathrm{tot}} \leq b_\nu + O(1)$. The same construction has $Q = N_{\mathrm{in}}$ and $C_{\mathrm{tot}} = O(\overline{N}_\nu)$ for every $0 < \nu \leq 1$. The exact worst-case query lower bound N_{in} remains valid up to $\nu = 1/2$. Without a supplied positive separation, deterministic verification has no uniform finite worst-case bit cap.

A deterministic verifier may need to inspect every input column. Checking only a few mixtures can leave a change hidden by cancellation, even though one column has a substantial total-variation difference. The lower bound already holds with two possible outputs, so increasing the output space does not remove this obstruction.

The supervisor does not need an exact real description of the worker. A quantized label at the promised resolution suffices, and its optimal length is proportional to the number $N_{\mathrm{in}}(N_{\mathrm{out}} - 1)$ of free probability coordinates times $\log(1/\nu)$. This guarantee depends on knowing a resolution at which a corruption must be visible.

Proof sketch. Basis queries reveal the columns, whose compact labels separate every promised unequal pair. For the query lower bound, use columns $(1/2, 1/2, 0, \ldots, 0)^\top$. Fewer than N_{in} query vectors leave a nonzero null direction; changing the first two rows in opposite directions keeps all columns normalized and hides a legal corruption. For the bit lower bound, pack many well-separated probability matrices and reduce verification on this packing to exact equality of finitely many labels. □

Proof. Construction and communication upper bound. The worker obtains every column using $q = e_j$, for $j \in [N_{\mathrm{in}}]$, and sends $\mathsf{Quant}_\nu(B)$. Alice computes her own label and compares them. If two probability columns share a label, then

$$\mathrm{TV}(p, p') \leq \sum_{a < N_{\mathrm{out}}} |p_a - p'_a| < \frac{N_{\mathrm{out}} - 1}{G_\nu} \leq \frac{\nu}{2}.$$

The first inequality uses the column-sum constraint to bound the last coordinate difference by the sum of the others. Thus a promised unequal pair cannot share the full matrix label, while equality always does. For $s = N_{\mathrm{out}} - 1 \geq 1$, $G_\nu + s \leq 4s/\nu$ when $\nu \leq 1$. Consequently

$$\log_2 \binom{G_\nu + s}{s} \leq s \log_2 \left(\frac{e(G_\nu + s)}{s} \right) \leq s \log_2(4e/\nu).$$

The ceiling in each column code adds at most N_{in} bits, giving $b_\nu \leq \overline{N}_\nu + N_{\mathrm{in}} = O(\overline{N}_\nu)$. Fixed start/return frames add $O(1)$ bits. For $\nu \leq 1/8$ this is $O(N_\nu)$.

Adaptive query lower bound. Let every column of the reference A^{c} equal $(1/2, 1/2, 0, \ldots, 0)^\top$. Follow a deterministic verifier on the honest input and collect its $m < N_{\mathrm{in}}$ query vectors in a matrix $\mathsf{Q} \in \mathbb{R}^{N_{\mathrm{in}} \times m}$. There is a nonzero row vector v with $v\mathsf{Q} = 0$; scale it so $\|v\|_\infty = \nu$. Define a worker matrix by

$$B_{1j} = 1/2 + v_j, \qquad B_{2j} = 1/2 - v_j, \qquad B_{aj} = 0 \quad (a > 2).$$

For $\nu \leq 1/2$ its columns are probability vectors and $\delta(A^{\mathrm{c}}, B) = \nu$. Yet $(B - A^{\mathrm{c}})\mathsf{Q} = 0$. Inductively every query response, subsequent query, and final decision matches the honest path. This contradicts correctness, proving the lower bound, including adaptivity.

Algorithm 5 Deterministic verification of probability matrices (Theorem E.1)

Input: Alice's private A, worker oracle B, known threshold ν.
Output: Exact decision under $B = A$ or $\delta(A, B) \geq \nu$.
1: Alice sends a fixed start frame.
2: The worker makes the N_{in} local probes $e_1, \ldots, e_{N_{\text{in}}}$.
3: Compute and send the b_ν-bit label $\mathsf{Quant}_\nu(B)$.
4: Alice accepts iff it equals $\mathsf{Quant}_\nu(A)$.
Queries: $Q = N_{\text{in}}$.
Total bits: $C_{\text{tot}} \leq b_\nu + O(1) = O(N_\nu)$ for $\nu \leq 1/8$.

A packing and the bit lower bound. Project each probability column onto its first $s = N_{\text{out}} - 1$ coordinates. Its simplex has volume $1/s!$. A ball of radius ν in total variation projects inside an ℓ_1 ball of radius 2ν, whose volume is $(4\nu)^s/s!$. Choose a maximal finite packing $\mathcal{P}_\nu$ of $\mathcal{X}_{\text{p}}$ at pairwise maximum-column distance greater than ν. Such a packing exists by compactness, and its closed radius-ν balls cover the space. Comparing volumes in the N_{in} projected columns gives, with $J_\nu = |\mathcal{P}_\nu|$,

$$J_\nu \geq (1/(4\nu))^{D_{\text{p}}}, \qquad \log_2 J_\nu \geq N_\nu/3 \quad (\nu \leq 1/8). \tag{47}$$

Every unequal pair from the packing is a valid alternative. Even giving the worker its matrix for free leaves the problem of exact equality on J_ν labels. Lemma B.2, with error zero, requires $\Omega(\log_2 J_\nu) = \Omega(N_\nu)$ bits. This lower bound holds regardless of the number of local probes. Without a supplied gap one can restrict to arbitrarily large finite sets of distinct probability matrices, with every unequal pair valid. The same exact equality bound diverges, proving the final assertion. □

E.2 Randomized Verification: Proof for Theorem 5.1

Theorem 5.1 is stated in Section 5. We prove its one-query construction, optimal communication bound, and conditions for joint attainment here. Algorithm 6 gives both the one-query hashing and exact-quantization protocols.

Supplementary guarantees. At zero error, the optimal private-coin query cap is N_{in} and the optimal total communication is $\Theta(N_\nu)$. For every $0 < \nu \leq 1$, including the larger-gap range, hashing gives $Q = 1$ and $C_{\text{tot}} = O(\log_2(\overline{N}_\nu + 1) + L)$. Without a supplied positive gap, no finite worst-case bit cap achieves private-coin error below $1/2$. These additional assertions are proved along with the theorem below.

Proof of Theorem 5.1 and its supplementary guarantees.
A finite rational query and quantitative separation. Let K be the smallest power of two at least $\max\{1, 4(N_{\text{in}} - 1)/\varepsilon\}$. For $j \in \{0, \ldots, K-1\}$ put

$$\begin{aligned} z_j &= 1 + j/K, \qquad s(z) = \sum_{b=0}^{N_{\text{in}}-1} z^b, \\ q(z) &= \frac{(1, z, \ldots, z^{N_{\text{in}}-1})^\top}{s(z)}, \qquad \tau = \frac{\nu}{N_{\text{out}} N_{\text{in}}^2\, 2^{N_{\text{in}}} (3K)^{N_{\text{in}}-1}}. \end{aligned} \tag{48}$$

Every query is a rational probability vector generated locally from its index. Alice privately selects a uniform index J. For a fixed promised unequal pair let $E = B - A$. Some column has ℓ_1 discrepancy at least 2ν, so some entry has magnitude at least $2\nu/N_{\text{out}}$. Choose its row $v = (v_1, \ldots, v_{N_{\text{in}}})$ and the polynomial $p(z) = \sum_{b=1}^{N_{\text{in}}} v_b z^{b-1}$.

At most $N_{\text{in}} - 1$ grid points can satisfy $|vq(z_j)| < \tau$. Otherwise take N_{in} such distinct points $x_1, \ldots, x_{N_{\text{in}}}$. Since $1 \leq x_i < 2$, $|p(x_i)| < N_{\text{in}} 2^{N_{\text{in}}-1} \tau$. Lagrange interpolation yields

$$p(z) = \sum_{i=1}^{N_{\text{in}}} p(x_i) \frac{\prod_{b \neq i}(z - x_b)}{\prod_{b \neq i}(x_i - x_b)}.$$

The sum of absolute coefficients of each numerator is at most $3^{N_{\text{in}}-1}$, and each denominator has magnitude at least $K^{-(N_{\text{in}}-1)}$. Thus every coefficient of p has magnitude strictly less than

$$N_{\text{in}}^2 2^{N_{\text{in}}-1}(3K)^{N_{\text{in}}-1}\tau = \nu/(2N_{\text{out}}),$$

contradicting the selected large coefficient. This also covers $N_{\text{in}} = 1$ using empty products equal to one. In particular,

$$\Pr_J\{\|Bq(z_J) - Aq(z_J)\|_\infty < \tau\} \le (N_{\text{in}} - 1)/K \le \varepsilon/4. \tag{49}$$

Finite response labels and private hashing. Quantize a response y to $\mathsf{Resp}_\tau(y) = (\lfloor 2y_a/\tau \rfloor)_{a\in[N_{\text{out}}]}$. Equal labels imply coordinate differences strictly below $\tau/2$, so the labels differ outside the event in (49). Each label has an injective fixed binary encoding of length

$$D_{\text{resp}} = N_{\text{out}} \left\lceil \log_2(1 + \lfloor 2/\tau \rfloor) \right\rceil .$$

Apply Lemma B.3 to these finite binary strings, with collision probability at most $\varepsilon/4$. Alice chooses its hash index independently of J and sends both indices to the worker. The worker makes its single local query and sends the selected $\lceil \log_2(16/\varepsilon) \rceil$-bit hash of its response label. Alice computes the corresponding label from $Aq(z_J)$ and compares. Conditional on each query index producing unequal labels, the independent hash collides with probability at most $\varepsilon/4$. Thus the total miss is at most $\varepsilon/2$, and equal matrices always pass.

The index of the query costs $\log_2 K = O(\log_2(N_{\text{in}} + 1) + L)$ bits. The displayed precision gives

$$D_{\text{resp}} = O(N_{\text{out}}[\log_2(N_{\text{out}}/\nu) + N_{\text{in}} \log_2(N_{\text{in}} + 1) + N_{\text{in}}L]) .$$

For $0 < \nu \le 1$, the quantity $\overline{N}_\nu$ bounds constant multiples of $N_{\text{out}}N_{\text{in}}$ and $N_{\text{out}} \log_2(1/\nu)$ from below. Hence

$$D_{\text{resp}} = O\big(\overline{N}_\nu[\log_2(\overline{N}_\nu + 1) + L + 1]\big) , \qquad \log_2(D_{\text{resp}} + 1) = O(\log_2(\overline{N}_\nu + 1) + L).$$

Lemma B.3 therefore charges $O(\log_2(\overline{N}_\nu + 1) + L)$ bits for its independent hash index and return value, as well as for the query index. Fixed framing adds $O(1)$. This proves the one-query upper bound, with all seeds paid. For $\nu \le 1/8$ it is $O(F_\nu(\varepsilon))$. Taking the better of this protocol and deterministic quantization gives the minimum bit order. If $L \le N_\nu$, then $F_\nu(\varepsilon) \le 2N_\nu$, since $\log_2(N_\nu + 1) \le N_\nu$ for $N_\nu \ge 1$; hashing is consequently within a constant of the minimum. If $N_{\text{in}} \ge 1/\varepsilon$, then $L \le \log_2 N_{\text{in}} \le N_\nu$, establishing the large-input simultaneous guarantee described in the main text.

Lower bounds and zero error. With no probes, neither party can obtain worker-matrix information, so the two-sided error is at least $1/2$ at any reference with an allowed alternative. Thus the positive-error optimal query cap is one. Restrict inputs to the packing (47) and let $b = C_{\text{tot}}$. The rectangle inequality of Lemma B.2 gives $\varepsilon + 1/J_\nu \ge c2^{-b}$, implying $\Omega(\min\{\log_2 J_\nu, L\})$ bits. A constant number of private repetitions reduces error to $1/64$, and its private packing inequality adds $\Omega(\log_2(\log_2 J_\nu + 1))$ bits. Their maximum is within an absolute factor of $\min\{N_\nu, \log_2(N_\nu + 1) + L\}$ by (47). The no-communication obstruction handles bounded parameters. This proves the matching bit lower bound without any restriction on the number of probes.

For zero error, consider the honest execution at the central reference A^{c} from the deterministic proof. If its hard query cap were less than N_{in}, select a halting history specified by finite prefixes of both private tapes that has positive probability. Such a history exists because there are countably many finite histories and termination is almost sure. Its query list has fewer than N_{in} vectors. The same nullspace construction gives a fixed legal alternative at distance ν producing the same history for every tape in that positive-probability cylinder. On the honest input the history must accept, so the alternative has positive error, a contradiction. Basis queries and quantization attain N_{in} probes and $O(N_\nu)$ bits. Setting the rectangle error to zero gives the matching bit lower bound even for randomized protocols.

Finally, without a supplied gap restrict inputs to an arbitrarily large finite set of distinct probability matrices. Every unequal pair is then valid. For any fixed error below $1/2$, a finite number of independent repetitions makes the error at most $1/64$. The private packing inequality would bound the set's size in terms of the fixed bit cap, a contradiction. This proves the gap-free impossibility; it does not rely on any ideal continuously distributed random query. □

Algorithm 6 Private randomized probability verification (Theorem 5.1)

Input: Alice's A, worker oracle B, known $0 < \nu \leq 1/8$, target $0 < \varepsilon \leq 1/3$; mode ONEQUERY or MINIMUMBITS.
Output: One-sided miss probability at most ε.
1: **if** mode is MINIMUMBITS and $N_\nu \leq F_\nu(\varepsilon)$ **then**
2: Run Algorithm 5 and return its decision.
3: **end if**
4: Compute the shared parameters K, τ and grid in (48).
5: Alice privately chooses uniform $J \in \{0, \ldots, K-1\}$ and an independent finite-table hash index with collision probability $\varepsilon/4$.
6: Send both indices in a charged start frame.
7: The worker generates $q(z_J)$, makes one local probe, and computes $\mathsf{Resp}_\tau(Bq(z_J))$.
8: Send its selected hash to Alice; she accepts iff it equals the hash of $\mathsf{Resp}_\tau(Aq(z_J))$.
Queries: hashing 1; exact branch N_{in}.
Total bits: hashing $O(F_\nu(\varepsilon))$; exact branch $O(N_\nu)$; MINIMUMBITS $O(\min\{N_\nu, F_\nu(\varepsilon)\})$.

F PROBABILISTIC SETTING: RESULTS FOR MULTIPLE WORKERS

F.1 DETERMINISTIC IDENTIFICATION

Theorem F.1 (Optimal deterministic probability identification)**.** *Let $t \geq 2$, $N_{\mathrm{out}} \geq 2$, $N_{\mathrm{in}} \geq 1$, and $0 < \nu \leq 1/8$. In the complete network, under the exactly-one corruption promise,*

$$Q^{\mathrm{det}}_{\mathrm{p},t}(\nu) = (t-1)N_{\mathrm{in}} \qquad \textit{number of queries}, \tag{50}$$

$$C^{\mathrm{det}}_{\mathrm{p},t}(\nu) = \Theta(tN_\nu) \qquad \textit{total communication bits}. \tag{51}$$

Algorithm 7 attains both simultaneously using $Q \leq (t-1)N_{\mathrm{in}}$ and $C_{\mathrm{tot}} \leq (t-1)b_\nu + O(t)$. For every $0 < \nu \leq 1$ it has the same query cap and $C_{\mathrm{tot}} = O(t\overline{N}_\nu)$. The exact deterministic query lower bound remains valid for every $\nu \leq 1/2$. Without a supplied positive gap, no uniform finite deterministic total-bit cap is possible.

The promise of exactly one corruption lets Alice test $t-1$ workers and infer the last one if all tests pass. Nevertheless, in the worst case each tested worker needs every input column checked. Worker-to-worker messages can move information already obtained, but cannot replace missing probes of a worker's stored matrix.

The communication cost counts all links. Complete connectivity does not improve the order tN_ν, since identifying a possibly different worker requires enough information across many worker cuts.

Proof sketch. Use the deterministic single-worker protocol along a fixed traversal of $t-1$ workers. For the lower bound, follow the all-honest execution. Every worker except its final fallback must have a separating query list, requiring N_{in} calls at the central reference. For communication, restrict inputs to the probability-matrix packing and apply the finite-family cut argument already proved for the discretized setting. Its proof depends only on the number of possible input labels, not on the arithmetic used by the oracle. □

Proof. Testing workers $1, \ldots, t-1$ with Algorithm 5 gives the upper bounds. Every honest worker passes, and every promised corrupted tested worker rejects. If all tested workers pass, the unique corruption is at t. The fixed traversal has $O(t)$ framing cost, counted in the theorem.

For the query lower bound, first centralize any peer protocol for analysis. A simulator knowing A can reproduce all faithful controller states and messages from the previous oracle answers. When a new local probe is requested, it makes exactly that addressed call. Thus peer communication does not reduce the possible number of probes. At the central reference A^{c}, run the simulated protocol on the all-equal tuple and let f be its final output. If fewer than $(t-1)N_{\mathrm{in}}$ probes are used, some worker $i \neq f$ has fewer than N_{in} queries on this path. The nullspace construction in the single-worker proof produces a matrix at distance ν invisible at i. Changing only that worker preserves the full transcript and the incorrect output f. This proves the lower bound and includes adaptive protocols and arbitrary peer messages.

Algorithm 7 Deterministic probability identification on a complete network (Theorem F.1)

Input: Alice's A, $t \geq 2$ faithful worker oracles, known ν.
Output: Correct corrupted-worker index on every valid input.
1: **for** $i = 1, \ldots, t-1$ in a fixed traversal of Alice–worker links **do**
2: Run Algorithm 5 with worker i.
3: If it rejects, return i.
4: **end for**
5: Return t.
Queries: $Q \leq (t-1)N_{\mathrm{in}}$.
Total bits: $C_{\mathrm{tot}} \leq (t-1)b_\nu + O(t) = O(tN_\nu)$ for $\nu \leq 1/8$.
Peer links remain available; every transmitted bit is included.

For completeness, the finite-family network lower bound in Appendix C.2.3 applies to any family of J possible objects. Give each controller its full object for the lower bound. On a cut isolating a uniformly random worker, put object y there and object x at Alice and every other worker. Accept equality when the output differs from that isolated index. On $x \neq y$ the promise forces rejection; on $x = y$ the mean equality acceptance is $1 - 1/t \geq 1/2$. The sum of worker-cut traffic is at most twice the total traffic. Truncating a cut simulation at a constant multiple of its mean cost preserves constant mean diagonal acceptance and gives a hard cap $O(C_{\mathrm{tot}}/t + 1)$. The rectangle bound therefore gives $1/J \geq c2^{-c'(C_{\mathrm{tot}}/t+1)}$. The participation argument in the same appendix supplies $\Omega(t)$ for bounded J, so $C_{\mathrm{tot}} = \Omega(t \log_2 J)$ throughout. All these reductions are valid under the charged cut-simulation and all-input resource conventions specified above. Take $J = J_\nu$ in (47) to obtain $\Omega(tN_\nu)$. Without a known gap one can instead take arbitrarily large finite families of distinct matrices, proving unbounded total bits. □

F.2 Randomized Identification: Proof for Theorem 5.2

Theorem 5.2 is stated in Section 5. Algorithm 8 implements its fixed-traversal hashing protocol and exact-quantization alternative. Its query-minimal mode attains the exact query optimum, separately from bit minimization, as follows.

Exact query minimization and supplementary guarantees. Throughout this subsection assume $N_{\mathrm{in}} \geq 2$. For $N_{\mathrm{in}} \geq 2$ define

$$h = \lfloor t(1-\varepsilon) \rfloor, \qquad \Delta = h + 1 - t(1-\varepsilon) > 0, \qquad \eta = \min\{1/4, \Delta/(2h)\}.$$

The query-minimal mode has $Q \leq h$ and $C_{\mathrm{tot}} = O(h[\log_2(t+1) + F_\nu(\eta)])$. At zero error, private coins do not improve either optimum: $Q^{\mathrm{rand}}_{\mathrm{p},t}(\nu, 0) = (t-1)N_{\mathrm{in}}$ and $C^{\mathrm{rand}}_{\mathrm{p},t}(\nu, 0) = \Theta(tN_\nu)$. Without a supplied positive gap, no uniform finite private total-bit cap achieves the required error bound. The proof below includes these supplementary claims.

Proof of Theorem 5.2 and its supplementary guarantees.
Simultaneous query and bit upper bounds. Run Algorithm 6 on workers $1, \ldots, t-1$, stopping at the first rejection and otherwise outputting t. In one-query mode this uses at most $t-1$ probes and $O(tF_\nu(\varepsilon))$ bits, including every copied seed and $O(t)$ framing. Only the sole corrupted worker can be missed; an error target ε per test suffices. The exact mode uses at most $(t-1)N_{\mathrm{in}}$ probes and $O(tN_\nu)$ bits. The single-worker branch rule and its $L \leq N_\nu$ argument prove the total-bit upper bound and show that both optimal orders are attained whenever $L \leq N_\nu$. The same constructions with $\overline{N}_\nu$ in place of N_ν give upper bounds for every $0 < \nu \leq 1$.

Query lower bound and the finite-coin boundary. Centralize peer communication as in the deterministic proof and follow the all-honest execution at A^{c}. Fix all tapes for analysis. If a cap c is imposed, at most c distinct workers are probed on that path. Place a fixed legal unequal matrix at a uniformly chosen worker. For an unprobed index the execution is unchanged, and at most one such index is correctly named. Thus, for each fixed matrix,

$$\Pr\{\text{correct index}\} \leq (c+1)/t, \qquad c \geq t(1-\varepsilon) - 1. \tag{52}$$

For $N_{\mathrm{in}} \geq 2$, this already implies $c \geq \lfloor t(1-\varepsilon) \rfloor$ unless $t(1-\varepsilon)$ is an integer. In that boundary case let $c = t(1-\varepsilon) - 1$. To attain the coverage equality, the all-honest execution must probe exactly c

distinct workers and output an unprobed fallback almost surely; otherwise the right side of (52) is strict for every fixed alternative. Since there are at most c probes, each of these workers has exactly one probe.

There is a positive-probability all-honest halting history fixed by finite prefixes of all private tapes. This uses the finite-fair-bit and all-input termination assumptions. Pick a probed worker i on this history. Since $N_{\mathrm{in}} \geq 2$, its single query has a nonzero real null direction. The central-reference perturbation in the deterministic proof gives a fixed B at distance ν invisible to this probe. If B is placed at worker i, every execution in that cylinder has the same transcript and the incorrect unprobed fallback. For this fixed B, at most c of the t possible corruption indices can be correct on that positive-probability cylinder, while at most $c+1$ can be correct on every other tape. This makes the averaged coverage inequality strict, a contradiction. Thus one additional probe is necessary at the integer boundary, proving the lower half of (17).

Achieving the exact query minimum with finite bits. For $N_{\mathrm{in}} \geq 2$, choose a uniform h-subset of workers and an independent uniform fallback from its complement, where h, Δ, η are defined above. Test each chosen worker with the one-query verifier at miss target η. Return a rejection index if found, and otherwise the fallback. For every fixed corrupted index and matrix its error is at most

$$\frac{t-h-1+h\eta}{t} \leq \frac{t-h-1+\Delta/2}{t} < \varepsilon.$$

The random subset and fallback are chosen locally at Alice. Sending explicit worker addresses costs $O(h \log_2(t+1))$ bits, and the h tests cost $O(hF_\nu(\eta))$. Thus the protocol has finite hard communication and attains the exact probe optimum. Uniform finite choices can be sampled exactly with almost-surely finite fair-bit computation; its local random-bit count is not communication.

Total communication lower bound. Restrict all inputs to the packing $\mathcal{P}_\nu$ of size J_ν from (47). The reduction and inequalities of Appendix C.2.3 depend only on this finite family and the charged communication convention. Here is their application. Across a cut isolating worker i, one side holds y and the other simulates Alice and the remaining workers on x. Accept equality if the output is not i. For unequal x, y this acceptance is at most ε. Averaging a uniform cut on equal inputs makes diagonal acceptance $1 - 1/t$ and mean cut traffic at most $2C_{\mathrm{tot}}/t$. Exposing the private tapes for this lower-bound calculation and truncating at a sufficiently large constant multiple of that mean gives, by Lemma B.2,

$$\varepsilon + 1/J_\nu \geq c_0\, 2^{-c_1(C_{\mathrm{tot}}/t+1)}$$

for absolute constants $c_0, c_1 > 0$. Together with the participation bound $C_{\mathrm{tot}} = \Omega(t)$ this yields $\Omega(t \min\{\log_2 J_\nu, L\})$ bits.

For the private dimension term, reduce error to $1/64$ using a constant number of independent repetitions and majority output. On a uniformly chosen all-equal packing input, let a_i be the mean probability of outputting i, and T_i the mean traffic across its cut. At least $t-1$ indices have $a_i \leq 3/4$. Fix each such cut, so its two sides retain independent private tapes. Its equality acceptance has mean at least $1/4$ on the diagonal. Truncating at $\max\{1, \lceil 16T_i \rceil\}$ crossing bits retains mean at least $3/16$ and does not increase off-diagonal acceptance. The private packing inequality in Lemma B.2 gives $T_i = \Omega(\log_2(\log_2 J_\nu + 1))$. Summing these cuts and using $\sum_i T_i \leq 2O(C_{\mathrm{tot}})$ proves the additional lower bound $\Omega(t \log_2(\log_2 J_\nu + 1))$. Bounded parameters are handled by participation. Combining the two lower bounds and $\log_2 J_\nu \geq N_\nu/3$ proves (18). This argument allows arbitrary peer traffic and arbitrary numbers of oracle calls.

Zero error and the absence of a supplied gap. For zero error and a cap below $(t-1)N_{\mathrm{in}}$, select a positive-probability all-honest finite-prefix history at A^{c}. Some worker different from its fallback has fewer than N_{in} probes. A fixed nullspace perturbation at that worker preserves the history and creates positive error, exactly as in the single-worker proof. Thus finite private coins cannot improve the deterministic query optimum at zero error. Setting the rectangle error to zero in the finite-family cut argument gives the exact bit lower bound $\Omega(tN_\nu)$, and deterministic quantization attains it. Finally, without a supplied gap use arbitrarily large finite families of distinct matrices in the private cut-packing argument. Its lower bound diverges with the family size, ruling out any uniform finite total-bit cap at error at most $1/3$. □

Algorithm 8 Private probability identification on a complete network (Theorem 5.2)

Input: Alice's A, $t \geq 2$ faithful workers with $N_{\text{in}} \geq 2$, known $0 < \nu \leq 1/8$, target $0 < \varepsilon \leq 1/3$; mode OnePerWorker, MinimumBits, or MinimumQueries.
Output: Corrupted-worker identification with error at most ε.

if mode is MinimumQueries **then**
  Set h by (17); choose a uniform h-subset S and a uniform fallback $f \in [t] \setminus S$ at Alice.
  Set $\eta = \min\{1/4, [h + 1 - t(1 - \varepsilon)]/(2h)\}$ and choose one-query tests at target η.
else
  Set $S = (1, \ldots, t - 1)$ and $f = t$.
  In OnePerWorker use one-query tests at target ε; in MinimumBits use the bit-branch rule of Algorithm 6.
end if
for each $i \in S$ in the selected order **do**
  Send the needed charged instructions and seeds over Alice's link; run the chosen single-worker test and return i if it rejects.
end for
Return f.

Queries: fixed hash $Q \leq t - 1$; exact $Q \leq (t - 1)N_{\text{in}}$.
Total bits: fixed hash $O(tF_\nu(\varepsilon))$; exact $O(tN_\nu)$.
Query-minimal mode: $Q \leq h$; bits $O(h[\log_2(t + 1) + F_\nu(\eta)])$.
All worker addresses, seed copies, frame boundaries, and control are charged.