# Separating Quantum Indistinguishability Obfuscation from Falsifiable Assumptions

Mohammed Barhoush[1], Tomoyuki Morimae[2], and Ramis Movassagh[1]

[1] IonQ Inc.
[2] Kyoto University

**Abstract.** Quantum indistinguishability obfuscation (qIO) aims to make a quantum circuit unintelligible while preserving its functionality. It serves as a foundational primitive for advanced applications, such as witness encryption (WE) for QMA, non-interactive zero-knowledge arguments for QMA, and attribute-based encryption for BQP. Despite its importance, constructing qIO from standard assumptions remains a major open problem.

In this work, we prove that the security of WE for QMA cannot be based on any falsifiable cryptographic assumption via a restricted class of quantum black-box reductions. Because qIO for null quantum circuits implies WE for QMA, this also separates null–qIO from falsifiable assumptions. Since almost all standard cryptographic assumptions are falsifiable, our result presents a barrier to basing qIO on standard cryptographic assumptions.

The reductions we rule out are restricted: the reduction must query the adversary classically, non-adaptively, at the same security parameter, and only on honestly generated ciphertexts. Moreover, our impossibility applies only to WE with classical ciphertexts, and therefore does not rule out qIO with obfuscators whose output is a quantum state. Ruling out more general reductions, as well as more general forms of WE and qIO, remains open.

Our impossibility relies on the existence of a QMA–QCIP[2] gap problem, an average-case assumption postulating a QMA language that cannot be verified with two messages of classical communication.

# 1 Introduction

## 1.1 Background

Program obfuscation aims to make code unintelligible while preserving its functionality. The strongest formalization, virtual black-box (VBB) obfuscation, is famously impossible for general circuits [7]. This impossibility led Barak et al. [7] to propose the weaker notion of *indistinguishability obfuscation* (IO), which only requires that obfuscations of two functionally equivalent circuits be computationally indistinguishable. Despite this seemingly modest guarantee, classical IO has proven to be a central hub of modern cryptography [22, 41]. A long line of work recently culminated in a construction of classical IO from well-founded assumptions [33]. However, one of the underlying assumptions (on bilinear maps) is broken by quantum attacks, and no post-quantum instantiation from well-founded assumptions is known.

As quantum computers mature, it is natural to ask for obfuscation of *quantum* programs. Beyond simply protecting quantum software, quantum *indistinguishability obfuscation (qIO)* acts as a master primitive for quantum-native cryptography. Even in its weakest useful form—obfuscation of *null* quantum circuits, which reject every input—Bartusek and Malavolta [13] demonstrated that qIO implies witness encryption (WE) for QMA, a publicly verifiable non-interactive zero-knowledge argument system for QMA, and attribute-based encryption for BQP.

In the QMA setting, the statement is classical (e.g., "this local Hamiltonian has a low-energy state"), and the witness is a quantum state. Witness encryption—first introduced for NP by Garg et al. [24]—for QMA is particularly relevant to this work so we give an informal definition here.

**Definition 1 (Witness Encryption for QMA (Informal)).** *Let $L \in \mathsf{QMA}$. A WE scheme* $(\mathsf{WE.Enc}, \mathsf{WE.Dec})$ *for $L$ consists of the following quantum polynomial time algorithms:*

- $\mathsf{WE.Enc}(1^n, x, m)$*: Takes as input a statement $x$, security parameter $n$, and message $m \in \{0,1\}$. It then outputs a classical ciphertext $ct$.*
- $\mathsf{WE.Dec}(x, ct, |w_x\rangle)$*: Takes as input a statement $x$, a classical ciphertext $ct$, and a quantum witness $|w_x\rangle$ of $x$. It then outputs a message $\widetilde{m} \in \{0,1\}$.*

WE allows anyone to encrypt a message by using a statement $x \in L$ such that any witness for $x$ decrypts it, while security requires that encryptions using a false statement $x \notin L$ computationally hide the message. The contrast with classical NP is instructive: WE for NP follows from classical IO [24, 41] and hence, via [33], from falsifiable assumptions. A falsifiable assumption [38, 26] is one that can be modeled as an interactive game between an efficient challenger and an adversary, where the challenger efficiently decides whether the adversary won (Definition 13). The vast majority of cryptographic assumptions—one-way functions, LWE, DDH, and so on—are falsifiable. Whether WE for QMA admits any foundation in falsifiable assumptions has remained completely open.

More generally, the state of qIO constructions is unsettled, primarily because they have yet to be realized in the plain model with a security proof from standard assumptions. For example, the null–qIO scheme of [13] is proven secure relative to a classical oracle. Other qIO variants face the same barrier: obfuscators for pseudo-deterministic quantum circuits [12, 10] and for unitary programs [32] are likewise proven secure only within an idealized classical-oracle model. Most recently, [11] constructed classical obfuscation of pseudo-deterministic quantum circuits, yet their security proof is again given only relative to a classical oracle. On the negative side, while quantum VBB obfuscation is known to be impossible [4, 1], these impossibility results do not apply to qIO.

In short, after a decade of sustained effort, we neither have a construction of qIO in the plain model with a security proof from standard assumptions, nor an explanation for its absence. This state of affairs motivates the central question of this work:

*Can we build qIO (or even just WE for* QMA*) from standard assumptions?*

### 1.2 Our Results

We provide a negative answer to this question for a natural class of security proofs. We introduce a new assumption, the existence of a QMA–QCIP[2] gap problem, and by assuming it, we prove that the security of WE for QMA cannot be based on any falsifiable cryptographic assumption via a restricted class of quantum black-box reductions. Since null–qIO implies WE for QMA, the same separation applies to null–qIO.

Note that WE and qIO are themselves non-falsifiable: verifying that an adversary has won the security game requires deciding whether its chosen statement lies outside the language, or whether the two quantum circuits it produced are both null, which cannot be done efficiently. Consequently, one cannot hope to rule out reductions from all assumptions: any primitive can be trivially based on the (non-falsifiable) assumption that the primitive itself is secure. Following Naor [38] and Gentry–Wichs [26], restricting attention to falsifiable assumptions is therefore the standard way to obtain meaningful impossibility results for non-falsifiable primitives [26, 8, 3].

The main result of this paper is the following:

**Theorem 2 (Main result, informal).** *Let $L$ be a* QMA *language admitting a QMA–QCIP[2] gap problem, and let $\Pi$ be a WE scheme for $L$. Then for every falsifiable assumption, either the assumption is false, or there is no quantum black-box reduction establishing the security of $\Pi$ from it, provided the queries made by the reduction algorithm to the adversary are security-parameter-preserving, non-adaptive, classical, and only honestly generated ciphertexts.*

Because null–qIO implies WE for QMA [13], we obtain the following corollary:

**Corollary 3 (Informal).** *Assume that there exists a* QMA *language admitting a QMA–QCIP[2] gap problem. Then for every falsifiable assumption, either the*

*assumption is false, or there is no quantum black-box reduction establishing the security of any null–qIO from it, subject to the same restrictions on the reduction as in Theorem 2.*

Note that Theorem 2 only covers WE with classical ciphertexts. Hence, Corollary 3 does not rule out obfuscators that output quantum states, such as the obfuscators for pseudo-deterministic quantum circuits [12, 10] and for unitary programs [32]. (See also Section 1.4).

Our separation assumes the existence of a QMA–QCIP[2] gap problem[3], defined as follows:

**Definition 4 (QMA–QCIP[2] gap problem, informal).** *A QMA–QCIP[2] gap problem for a language $L \in$ QMA consists of a distribution over yes-instances that is efficiently samplable together with the (quantum) witnesses, and a distribution over no-instances that may be sampled inefficiently, such that no efficient quantum algorithm can distinguish the two distributions, even when given classical access to an oracle that decides every two-message interactive proof over a classical channel.*

Informally, this assumption is an average-case formulation of the statement that some QMA languages cannot be verified via two messages of classical communication. This is well supported by the current cryptographic landscape, primarily stemming from the known barriers to achieving two-message classical verification of QMA. Mahadev's breakthrough classical verification of quantum computation protocol [36] requires four messages. While a long line of subsequent works has refined and optimized various aspects of her original scheme [27, 46, 39, 6, 2, 18, 9], four messages remain the state of the art in the plain model [4]

Another circumstantial support is the recent black-box separation by Barhoush, Morimae, Nishimaki, and Yamakawa [8], which proves that non-interactive classical verification of QMA (technically, a two-message protocol with a one-message setup) cannot be based on falsifiable assumptions. This provides a negative result reinforcing the conjecture that QMA $\not\subseteq$ QCIP[2].[5] Our notion of QMA–QCIP[2] gap problem is directly inspired by the QMA–QCMA gap problem introduced there; however, the choice to strengthen the assumption by replacing QCMA with QCIP[2] in the gap problem is motivated by the interactive nature of the WE primitive, as we detail in Section 2.

Ultimately, our theorem should be read as a structural barrier: if one believes a falsifiable assumption is true, then no security proof of the stated form can exist to base WE for QMA on it. The four constraints on the reduction—non-adaptive access, preservation of the security parameter, classical queries, and honestly

[3] QCIP is the complexity class for interactive protocols over classical channels [16, 29], and QCIP[2] is its two-message version.

[4] Although a non-interactive scheme has been achieved in the random oracle model [2], it does not apply to the plain model.

[5] The result [8] is for classical verifiers, but the result can be actually extended to the quantum verifier case.

generated ciphertext queries (Definitions 19 and 20)—are natural restrictions discussed at length in Sections 1.4 and 2.

Finally, two features distinguish our separation from prior works in the simulatable-adversary paradigm [26, 3, 8]. First, ours is the first separation in this paradigm to apply to an interactive security game: the previous ones all target a one-shot proof system, in which the adversary's attack consists of producing a single message. Meta-reduction-based impossibilities do address interactive security notions, but they concern different assumption classes and impose different structural restrictions on the scheme and on the reduction; see Section 1.3. We hope our techniques open the door to separating further interactive security notions from all falsifiable assumptions, where one-shot techniques break down. Second, we rule out reductions even to falsifiable assumptions that are secure only against *uniform* adversaries. Previous separations [26, 3, 8] are proven for assumptions formulated against non-uniform adversaries, and thus say nothing about assumptions whose security is inherently uniform. An example of such an assumption is a keyless collision-resistant hash function, which can only be secure against uniform adversaries, since a non-uniform adversary can simply hardwire a fixed collision as advice.

### 1.3 Related Works

*Separations from falsifiable assumptions.* Gentry and Wichs [26] proved that (adaptively-sound) succinct non-interactive arguments (SNARGs) admit no black-box reduction to any falsifiable assumption (under subexponential hardness of some subset-membership problems), introducing the simulatable-adversary paradigm that underlies all subsequent work in this line, including ours. Alagic, Dachman-Soled, Shingane, and Struck [3] extended the SNARG separation to *quantum* black-box reductions. Closest to our work, Barhoush, Morimae, Nishimaki, and Yamakawa [8] separated non-interactive classical verification of QMA from all falsifiable assumptions, assuming a QMA–QCMA gap problem. Their result concerns a two-message argument system, where the adversary's attack consists of producing a single classical message. Ours concerns an encryption scheme, whose security game is interactive with two classical messages produced by the adversary. This difference drives most of the technical divergence between the two works (see Section 2).

*Meta-reductions.* A complementary technique for ruling out reductions for interactive security notions is the meta-reduction paradigm [21, 40], which turns the reduction itself into an efficient attack on the assumption, typically by restricting its rewinding behavior. Because the attack produced by a meta-reduction uses no advice, the technique also applies to assumptions that are secure only against uniform adversaries. On the other hand, the known meta-reduction impossibilities either concern restricted assumption classes, such as non-interactive assumptions [21], or rely on structural properties of the scheme under attack, such as special soundness and unique witnesses [40]. The simulatable-adversary paradigm instead exhibits an inefficient adversary that is indistinguishable from

an efficient simulator under the reduction's query access, and it applies to arbitrary falsifiable assumptions. The two approaches impose different restrictions and apply to different regimes.

*Negative results on witness encryption.* The only prior negative result targeting witness encryption itself concerns a stronger variant. Garg, Gentry, Halevi, and Wichs [23] show that *extractable* WE for NP with auxiliary input is implausible, under a special-purpose obfuscation conjecture. Their result rules out the primitive (conditionally), whereas ours concerns plain WE for QMA and rules out a class of security proofs from falsifiable assumptions. The two results are complementary, and neither subsumes the other.

*Constructions of WE from non-falsifiable assumptions.* Witness encryption for NP and null–IO for classical circuits can be built from evasive LWE [44, 43], a plausibly post-quantum assumption. Evasive LWE is non-falsifiable: its statement quantifies over auxiliary-input samplers and admits no efficient challenger.

*Black-box separations in quantum cryptography.* A growing body of work maps the black-box landscape of quantum primitives: oracle separations between pseudorandom unitaries and one-way functions [34], quantum extensions of classical separations [31, 5, 19], and separations among quantum primitives [20, 17]. These results separate primitives from primitives. In contrast, we separate a primitive from an entire class of falsifiable assumptions.

*Impossibility of black-box obfuscation.* VBB obfuscation is known to be impossible for general classical circuits against classical adversaries [7]. Alagic and Fefferman [4] initiated the study of quantum obfuscation and proved, assuming quantum-secure one-way functions, that quantum circuits cannot be VBB obfuscated into quantum states when the adversary may hold multiple copies of the obfuscated state. Subsequently, Alagic, Brakerski, Dulek, and Schaffner [1] showed that, under the quantum hardness of LWE, even classical circuits cannot be VBB obfuscated into quantum states, where a single copy of the obfuscated state suffices for the impossibility. Negative results are also known for indistinguishability obfuscation of classical circuits. Mahmoody, Mohammed, Nematihaji, Pass, and Shelat [37] rule out fully black-box constructions of IO from one-way functions and from collision-resistant hashing, assuming one-way functions exist and $\mathsf{NP} \not\subseteq \mathsf{coAM}$, and Goldwasser and Rothblum [28] show that statistically secure IO for all circuits would collapse the polynomial hierarchy. These results separate IO from specific primitives or rule out statistical security altogether, whereas we separate quantum IO from an entire class of assumptions.

### 1.4 Limitations and Open Problems

*Restrictions on the reduction.* Our impossibility result applies to reductions that (1) preserve the security parameter, (2) access the adversary non-adaptively and classically, and (3) query the adversary only on honestly generated ciphertexts.

We call such reductions *restricted WE–reductions* (Section 3.6). Restriction (3) reflects the intuition that a valid attack on WE is the ability to break honestly generated ciphertexts i.e. a ciphertext that was generated by running the honest encryption algorithm on a uniformly random bit. Hence, a reduction ought to be able to leverage this ability alone to attack the falsifiable assumption. However, removing any restriction is the main open problem left by this work. As a first step toward a full separation of WE for QMA from falsifiable assumptions, we believe these restrictions are reasonable.

Prior works [26, 3, 8] can lift the parameter-preserving, classical-access, and non-adaptivity restrictions by hardwiring the adversary's behavior as non-uniform advice. As we explain in Section 2, this route is unavailable in our interactive setting. In exchange, our proof dispenses with advice altogether: it rules out reductions even to falsifiable assumptions whose security holds only against uniform adversaries, which is unclear for the prior separations. We consider this a reasonable trade-off, but lifting either restriction would be interesting.

There is a further consequence of our security formulation. Since the adversary produces the instance itself, the reduction can never query the adversary on an instance of its own choosing. As in previous separations in this line of work [26, 3, 8], our separation thus concerns an adaptive-style security notion, and does not address reductions that establish selective, per-instance security.

*IO for classical circuits.* Our techniques do not, and provably cannot, extend to IO for classical circuits. The reason is structural: IO for classical circuits implies WE for NP, and there is no analog of the QMA–QCIP[2] gap problem for NP—the classical witness is a one-message certificate, so a QCIP[2] oracle decides the language outright. In fact, classical IO (without quantum security) can be constructed from falsifiable assumptions [33], and hence so can WE for NP, so no separation can exist for the classical notion unless one of the underlying assumptions is false.

A stronger notion is post-quantum IO, namely IO for classical circuits that is quantumly secure. A separation for post-quantum null–IO would face the same structural barrier, since null–IO for classical circuits gives WE for NP. Here, interestingly, no construction from well-studied falsifiable assumptions is known either. The known plausibly post-quantum constructions of WE for NP and null–IO rely on evasive LWE [44, 43], a non-falsifiable assumption, or on new LWE-with-hints assumptions introduced for IO [15, 25, 45, 30], which are falsifiable but ad hoc, and several of which have been broken [30]. Post-quantum null–IO for classical circuits is thus a curious intermediate case: neither a construction from standard assumptions nor a black-box separation is known, and our technique provably cannot supply the latter.

The status of post-quantum IO for *general* classical circuits is open in both directions: neither a construction from standard assumptions nor a black-box separation is currently known. This is a particularly important direction for future research, as general-circuit post-quantum IO serves as the primary known route to primitives such as public-key quantum money and quantum software copy-protection. While our structural barrier implies that proving such a sepa-

ration would require techniques different from ours, we hope that the ideas in this work help in this direction.

*qIO with quantum output.* Notably, we only consider obfuscators that output classical bit strings. Consequently, our impossibility leaves untouched any obfuscator whose output is a *quantum* state. This gap is not vacuous, as schemes of precisely this excluded kind exist. The obfuscators of Bartusek, Kitagawa, Nishimaki, and Yamakawa [12] and of Bartusek, Brakerski, and Vaikuntanathan [10] act on pseudo-deterministic quantum circuits and yield obfuscated programs encoded as quantum states, while Huang and Tang [32] obfuscate unitary programs, again into quantum states. Because each of these produces an obfuscation with quantum output, all three fall beyond the syntactic class considered here, and our result carries no implication for them.

This exclusion is structural, not merely an artifact of our proof. With quantum ciphertexts, we would need an oracle for two-message protocols over a *quantum* channel, that is, a $\mathsf{QIP}[2]$-type oracle. Since a quantum witness is itself a one-message quantum certificate ($\mathsf{QMA} = \mathsf{QIP}[1] \subseteq \mathsf{QIP}[2]$), such an oracle decides the language outright, so no gap problem relative to it can exist. This mirrors the barrier for $\mathsf{NP}$ above.

*Instantiating the assumption.* Our separation assumes the existence of a $\mathsf{QMA}$–$\mathsf{QCIP}[2]$ gap problem. It would be valuable to base this gap problem on concrete quantum hardness assumptions, or at least to give an oracle separation between $\mathsf{QMA}$ and $\mathsf{QCIP}[2]$.

The corresponding one-message question was recently settled: Bostanci, Haferkamp, Nirkhe, and Zhandry [14] give a classical oracle separating $\mathsf{QMA}$ from $\mathsf{QCMA}$. This makes the oracle separation of $\mathsf{QMA}$ from $\mathsf{QCIP}[2]$ the concrete next step, and it would be interesting to see whether their oracle already defeats two-message classical-channel protocols. Exploring this gap more concretely could weaken our assumption and strengthen the separation, and would be of independent interest for quantum complexity theory.

## 2 Technical Overview

### 2.1 The Simulatable-Adversary Paradigm

Our impossibility result follows the simulatable-adversary paradigm of Gentry and Wichs [26]. To rule out black-box reductions, one exhibits two adversaries against the primitive: a computationally *unbounded* adversary $\overline{P}$ that genuinely breaks it, and an *efficient* simulator $P$ that no efficient algorithm can distinguish from $\overline{P}$ under the query access the reduction is permitted.

Why does such a pair suffice? Suppose a reduction $\Sigma$ bases the security of $\mathsf{WE}$ on a falsifiable assumption represented by a $\mathsf{QPT}$ challenger $C$. Informally, a falsifiable assumption is defined by an efficient game in which no efficient adversary should be able to make $C$ output 1 with noticeable advantage. A black-box reduction $\Sigma$ is an algorithm that, given oracle access to an adversary breaking

$\mathsf{WE}$, can break the falsifiable assumption. Therefore, if $\overline{P}$ is a computationally unbounded adversary that breaks $\mathsf{WE}$, then $C$ outputs 1 with non-negligible advantage when interacting with $\Sigma^{\overline{P}}$. But $\Sigma$ is efficient, so replacing its oracle access $\overline{P}$ by $P$ shifts its winning probability by at most a negligible amount. In other words, $\Sigma^{P}$ still wins the falsifiable assumption game with non-negligible advantage, and it is efficient, since $P$ is efficient. Therefore, this shows that the assumption was already false since it can be broken with an efficient attack. In summary, either there does not exist such a reduction or the assumption is false. The entire separation therefore reduces to two tasks: constructing the pair $(\overline{P}, P)$, and proving that no permitted reduction tells them apart.

However, there are multiple challenges that come up when we try to apply the simulatable-adversary paradigm to our setting.

### 2.2 Technical Challenges

*Challenge 1: $\mathsf{WE}$ security game is not appropriate.* Before applying the simulatable-adversary paradigm, we first need to modify the $\mathsf{WE}$ security game. This is because the textbook definition of $\mathsf{WE}$ does not have a fundamental property of previous primitives that were separated using the simulatable-adversary paradigm. In particular, the adversary in previous separations [26, 8] gets to *choose* the instance $x$ as part of the attack, which turns out to be crucial. We formulate a security notion for $\mathsf{WE}$ with this property. Our $\mathsf{WE}$ security game (Definition 16) is *interactive*: rather than fixing an instance $x \notin L$ in advance and asking the adversary to distinguish encryptions of 0 and 1 under it, we let the adversary *choose* its own instance. It outputs an instance $x$, receives a ciphertext encrypted under $x$, and wins if it guesses the encrypted bit correctly while ensuring $x \notin L$ (see Section 3.5 for more details). Because the adversary now carries the extra burden of sampling the no-instance itself, this notion is weaker than the standard one, and ruling out reductions for a weaker notion yields a *stronger* impossibility result.

The price of this design is interactivity: the attack is no longer a static map but a two-phase process—the adversary first commits to an instance, and only then answers a ciphertext. This structure breaks the separation techniques of [26, 3, 8], which lean throughout on the non-interactive, one-shot nature of their adversaries, as we discuss in the next two challenges.

*Challenge 2: Finding $\overline{P}$.* The paradigm requires an unbounded adversary $\overline{P}$ that breaks the scheme while remaining indistinguishable from an efficient $P$. To show this indistinguishability, our strategy is to *reproduce the behavior of $\overline{P}$ and $P$ using a $\mathsf{QCIP}[2]$ oracle*. Here, a $\mathsf{QCIP}[2]$ oracle $O$ is defined as follows: For $\mathsf{QPT}$ algorithms, $V_1$ and $V_2$, and a string $x$, we define

$$\mathrm{val}_{V_1,V_2}(x) \;:=\; \max_{\{z_y\}_y} \Pr_{(\rho,y)\leftarrow V_1(x)} \big[V_2(x,\rho,y,z_y)=1\big], \tag{1}$$

where $\rho$ is a quantum state. Then a $\mathsf{QCIP}[2]$ oracle $O$ is defined by

$$O(V_1, V_2, x) = \begin{cases} 1 & \text{if } \mathrm{val}_{V_1,V_2}(x) \geq 2/3, \\ 0 & \text{if } \mathrm{val}_{V_1,V_2}(x) \leq 1/3. \end{cases} \tag{2}$$

On inputs with $1/3 < \mathrm{val}_{V_1,V_2}(x) < 2/3$ the output of $O$ is unconstrained. In other words, a $\mathsf{QCIP}[2]$ oracle can find the optimal prover of the $\mathsf{QCIP}[2]$ protocol specified by the verifier $(V_1, V_2)$, and decide whether the optimal prover makes the verifier $V_2$ accept with high probability or low probability.

The main difficulty is that it is unclear how to use a $\mathsf{QCIP}[2]$ oracle to simulate $P$ or $\overline{P}$. A $\mathsf{QCIP}[2]$ query returns the value achieved by an *optimal* prover depending on the reduction and will not in general act as $P$ or $\overline{P}$.

Previous works avoid this issue by taking a union bound over all possible prover strategies; see [26, 8, 3]. The details of their argument are not important but essentially, in their non-interactive setting, a prover gives only a single polynomial-length answer. Hence, there are at most singly exponentially many strategies. In our setting, the prover must answer every possible ciphertext, so there are $2^{2^{\mathsf{poly}(n)}}$ such functions. This space is too large for the same arguments used in previous works.

We are therefore forced to take a different approach: we construct $\overline{P}$ explicitly, requiring it to simultaneously satisfy two conditions:

1. it breaks the $\mathsf{WE}$ scheme; and
2. it can be simulated by a $\mathsf{QCIP}[2]$ oracle.

While achieving either property independently is straightforward, satisfying both simultaneously is the difficult part.

Our primary candidate is the *Bayesian decryptor*: given a ciphertext, it outputs the most likely plaintext. This satisfies Condition 1 optimally, as no strategy can achieve higher inversion accuracy.

However, it is not clear if this satisfies Condition 2. At first glance, there is no guarantee that the interaction between the reduction and the Bayesian decryptor can be simulated by a $\mathsf{QCIP}[2]$ oracle. The $\mathsf{QCIP}[2]$ oracle can be used to check for the existence of a prover that makes the verifier accept with $2/3$ probability. But this optimal prover has no inherent incentive to invert ciphertexts; it will simply adopt whatever behavior maximizes acceptance for the reduction.

We resolve this by redesigning the protocol to force the prover's optimal strategy to be inversion. Among the ciphertexts produced by the reduction, we interleave additional *planted ciphertexts* that we encrypt ourselves (and whose plaintexts we know). The verifier is instructed to reject instantly if any planted ciphertext is answered incorrectly. Under this mechanism, a prover cannot risk deviating from inversion: to pass the test, it must invert the planted ciphertexts, and because the planted ciphertexts are indistinguishable from the reduction's ciphertexts (by our assumption of honestly generated ciphertexts), it must invert all of them. Optimal play is thus forced to be Bayesian inversion, ensuring that the $\mathsf{QCIP}[2]$ oracle behaves exactly as $\overline{P}$. This requirement also highlights why

the reduction must produce honestly generated ciphertexts: planted ciphertexts can only remain hidden if drawn from the exact same distribution.

*Challenge 3: Our restrictions and advice.* Our separation is also limited to reductions that query the adversary on the same security parameter, non-adaptively and classically. In general, reductions may query their adversary on security parameters smaller than their own, adaptively, or quantumly. Prior separations [26, 3, 8] handle these situations by a sequence of hybrids that progressively strengthen the indistinguishability of $P$ and $\overline{P}$: from a single sample, to polynomially many samples, to adaptive quantum queries, and finally to adaptive quantum queries across all security parameters. Each of these steps exploits the fact that the adversary's behavior can be hardwired as non-uniform advice. However, in our setting, the adversary is an interactive map of exponential size, which prevents it from being compressed into non-uniform advice.

We therefore restrict our reductions in such a way that the queries are security-parameter preserving, non-adaptive, classical, and honestly-generated ciphertexts. Concretely, our restricted reduction $\Sigma$ interacts with its adversary as follows, for some polynomials $s$ and $\ell$ determined by the reduction. First, $\Sigma$ queries the first stage of the adversary $s$ times, receiving instances $x_1, \ldots, x_s$. Then, for each $i \in [s]$, $\ell$ ciphertexts $ct_{i,1}, \ldots, ct_{i,\ell}$ are generated by honestly running the encryption algorithm on $x_i$ and independently sampled uniform plaintext bits. Finally, $\Sigma$ queries the second stage of the adversary on the pairs $(x_i, ct_{i,j})$ and receives one plaintext guess for each. Thus, the queries are security-parameter-preserving, classical, non-adaptive, and made only on honestly generated ciphertexts.

However, there is a compensating benefit for our approach. As mentioned, the proofs in [26, 3, 8] rely on non-uniform advice across multiple steps, leaving it unclear whether their separations hold in a fully uniform setting. Our separation avoids this dependence, and therefore rules out black-box reductions even from falsifiable assumptions that are only uniformly secure.

### 2.3 Our Separation

Our separation is demonstrated in the following steps.

*Decryptability transfers to no-instances.* Nothing in the definition of $\mathsf{WE}$ prevents $\mathsf{WE.Enc}(1^n, x, \cdot)$ from becoming statistically lossy when $x \notin L$; in fact, lossiness on no-instances would naturally provide unconditional security. Our first main step (Lemma 30) rules out this possibility by showing that, assuming the existence of a $\mathsf{QMA}$–$\mathsf{QCIP}[2]$ problem, ciphertexts generated with no-instances can be (inefficiently) decrypted and the plaintext message recovered.

This is the first instance where we can employ the $\mathsf{QCIP}[2]$ oracle using the following key observation:

*"My ciphertext was decrypted correctly" can be expressed as a two-message protocol over a classical channel.*

To see this, consider a verifier that chooses $k$ random bits, encrypts each bit under an instance $x$, and sends the resulting ciphertexts to a prover. The prover returns one guess for each ciphertext, and the verifier checks these guesses against the original bits.

When $x \in L$, correctness of the witness encryption scheme guarantees the existence of a prover that succeeds with high probability. Therefore, when we query the $\mathsf{QCIP}[2]$ oracle with this verifier encrypting with a yes-instance, the oracle will return 1 since there exists a strategy (namely running the decryption algorithm) that will make the verifier accept. The $\mathsf{QMA}$–$\mathsf{QCIP}[2]$ gap assumption tells us that a $\mathsf{QCIP}[2]$ oracle cannot be used to distinguish yes- and no-instances of a language so the same decryption property must hold for almost all sampled no-instances. Therefore, even when $x \notin L$, there exists a (possibly inefficient) strategy that decrypts ciphertexts correctly with high probability.

*Choosing $P$ and $\overline{P}$.* Following the paradigm in Section 2.1, the problem reduces to defining two adversaries and proving their indistinguishability. Both are *two-stage* algorithms matching the interface of the security game: the first stage produces an instance and the second stage answers a ciphertext. Concretely, we write $\overline{P} = (\overline{P}_1, \overline{P}_2)$ for the unbounded adversary that breaks the witness encryption scheme $\Pi = (\mathsf{WE.Enc}, \mathsf{WE.Dec})$, and $P = (P_1, P_2)$ for the efficient adversary that no reduction can distinguish from it. Let $L$ be a $\mathsf{QMA}$ language that admits a $\mathsf{QMA}$–$\mathsf{QCIP}[2]$ gap problem. We define the two algorithms as follows:

- **Algorithm of $\overline{P}$.** The first stage samples an instance *outside* the language,

$$x \leftarrow \overline{P}_1(1^n), \qquad x \notin L,$$

  and the second stage is the Bayesian decryptor: it answers a ciphertext $ct$ with its most likely plaintext,

$$\overline{P}_2(x, ct) \;=\; \arg\max_m \Pr[\mathsf{WE.Enc}(1^n, x, m) = ct].$$

  Both stages may be computationally inefficient.
- **Algorithm of $P$.** The first stage samples an instance *inside* the language together with a quantum witness,

$$(x, |w_x\rangle^{\otimes t}) \leftarrow P_1(1^n), \qquad x \in L,$$

  and the second stage decrypts a ciphertext $ct$ using the honest decryption algorithm with this witness,

$$P_2(x, ct, |w_x\rangle^{\otimes t}) \;=\; \mathsf{WE.Dec}(x, ct, |w_x\rangle^{\otimes t}).$$

  Both stages are efficient, where the efficiency of $P_1$ is assumed as part of the $\mathsf{QMA}$–$\mathsf{QCIP}[2]$ gap problem.

Crucially, $\overline{P}_2$ is a well-defined, deterministic function: we established that ciphertexts generated from no-instances can still be decrypted correctly. This is what makes $\overline{P}$ a fixed adversary, independent of the reduction—and it is precisely this independence that lets us estimate its behavior with a $\mathsf{QCIP}[2]$ oracle, as we describe next.

*The Simulator Protocol and Hidden Tests.* For any algorithm $Q$, let

$$\Gamma^Q := \langle C, \Sigma^Q \rangle \tag{3}$$

denote the algorithm obtained by viewing the interaction between the falsifiable assumption challenger $C$ and the reduction $\Sigma^Q$ as a single algorithm. The output of $\Gamma$ is the output of $C$. We now build a $\mathsf{QPT}$ algorithm $\tilde{\Gamma}$ that, using only a $\mathsf{QCIP}[2]$ oracle, reproduces $\Gamma$'s output on both sides:

$$\left| \Pr\left[\Gamma^P(1^n) = 1\right] - \Pr_{x \leftarrow L_n}\left[\tilde{\Gamma}^{\mathsf{QCIP}[2]}(x) = 1\right] \right| \;\leq\; \epsilon(n), \tag{4}$$

$$\left| \Pr\left[\Gamma^{\overline{P}}(1^n) = 1\right] - \Pr_{x \leftarrow \overline{L}_n}\left[\tilde{\Gamma}^{\mathsf{QCIP}[2]}(x) = 1\right] \right| \;\leq\; \epsilon(n). \tag{5}$$

Here, $L_n := L \cap \{0,1\}^n$ and $\overline{L}_n := \{0,1\}^n \setminus L_n$. These equations are essentially saying that $\tilde{\Gamma}^{\mathsf{QCIP}[2]}(x)$ simulates $\Gamma^P$ when $x \in L$ and simulates $\Gamma^{\overline{P}}$ when $x \notin L$. This will allow us to later argue that both interactions are indistinguishable using $\mathsf{QMA}$–$\mathsf{QCIP}[2]$ gap problem.

Recall that the reduction $\Sigma$ has two stages. In the first stage, it queries the first adversary to get instances $x_1, \ldots, x_s$. Then $\Sigma$ generates ciphertexts under these instances, and queries the second adversary to get one plaintext guess for each ciphertext. We simulate these two stages by a $\mathsf{QCIP}[2]$ oracle.

The algorithm $\widetilde{\Gamma}^{\mathsf{QCIP}[2]}(x)$ queries the $\mathsf{QCIP}[2]$ oracle on $(V_1, V_2, x)$, which is defined as follows:

1. $V_1$ runs the interaction between the challenger $C$ and the reduction $\Sigma$, using the given instances $x = (x_1, \ldots, x_s)$ in place of the instances that $\Sigma$ would normally receive from the first stage of its $\mathsf{WE}$ adversary. It continues until $\Sigma$ has generated all of its ciphertext queries $ct_{i,j}$. We call these the *real* ciphertexts.
2. For each instance $x_i$, $V_1$ generates additional *test* ciphertexts: $k/2$ encryptions of 0 and $k/2$ encryptions of 1 for some polynomial $k$.
3. $V_1$ randomly shuffles the real and test ciphertexts and sends the resulting batch to the prover.
4. The prover returns one plaintext guess for each ciphertext. The verifier $V_2$ rejects if any test ciphertext is answered incorrectly. Otherwise, it extracts the answers corresponding to the real ciphertexts, restores their original order, and gives them to $\Sigma$ as the answers it would have received from the second stage of its $\mathsf{WE}$ adversary. The interaction between $\Sigma$ and $C$ then continues, and $V_2$ outputs the final output bit of $C$.

By definition, the $\mathsf{QCIP}[2]$ oracle evaluates this protocol against an optimal prover, namely one that maximizes the probability that $V_2$ outputs 1. We show that the hidden tests force such a prover to behave essentially like a decryptor.

Notice that the prover does not know which ciphertexts are real and which are tests. This hiding property follows from the restriction on the reduction. Every real ciphertext is an honest encryption under $x$, and every test ciphertext

is generated in exactly the same way. The random permutation then hides the positions of the tests. Within the set of ciphertexts encrypting a fixed bit, the prover cannot identify which ones came from the reduction and which ones were planted by the verifier. Therefore, deliberately answering a real ciphertext incorrectly creates a significant risk of answering a hidden test incorrectly and being rejected.

The hidden tests therefore force an optimal prover to decrypt all ciphertexts. As a result, $\tilde{\Gamma}^{\mathsf{QCIP}[2]}(x)$ approximates the interaction between $\Gamma$ and the appropriate decryptor: $P$ when $x \in L$, and $\overline{P}$ when $x \notin L$. This gives Eqs. (4) and (5).

We now apply the gap assumption to the $\mathsf{QPT}$ distinguisher $\tilde{\Gamma}$. The assumption states that no efficient algorithm can distinguish elements inside and outside the language, even given a $\mathsf{QCIP}[2]$ oracle. This means that the outputs of $\tilde{\Gamma}$ on the two instance distributions are computationally indistinguishable:

$$\left| \Pr_{x \leftarrow \overline{L}_n} \left[ \tilde{\Gamma}^{\mathsf{QCIP}[2]}(x) = 1 \right] - \Pr_{x \leftarrow L_n} \left[ \tilde{\Gamma}^{\mathsf{QCIP}[2]}(x) = 1 \right] \right| \leq \operatorname{negl}(n). \tag{6}$$

Chaining (4), (5), and (6) through the triangle inequality yields

$$\left| \Pr \left[ \Gamma^{P}(1^n) = 1 \right] - \Pr \left[ \Gamma^{\overline{P}}(1^n) = 1 \right] \right| \leq 3\epsilon(n). \tag{7}$$

*Conclusion.* Suppose a reduction $\Sigma$ bases the security of the $\mathsf{WE}$ scheme on a falsifiable assumption. Interacting with $\Sigma^{\overline{P}}$, the challenger in the falsifiable assumption game accepts with noticeable advantage because $\overline{P}$ breaks the $\mathsf{WE}$ scheme. By Eq. (7), replacing $\overline{P}$ with $P$ alters this acceptance probability by only a negligible amount.

However, $P$ is efficient. Therefore, $\Sigma^{P}$ is an efficient attack against the falsifiable assumption. It follows that the assumption must be false. Consequently, either no reduction of the stated form exists, or the underlying falsifiable assumption is false. This proves our main result, Theorem 2.

## 3 Preliminaries

### 3.1 Notation

In this work, we denote sets and algorithms by capital letters such as $X, A$ and distributions by bold such as $\mathbf{L}, \mathbf{E}$. All sets and distributions are assumed to be parameterized by some integer $n$, typically a security parameter. We write $x \leftarrow \mathbf{D}$ to say that $x$ is sampled according to the distribution $\mathbf{D}$. If $X$ is a set, then $x \leftarrow X$ means that $x$ is chosen uniformly at random from the set. Let $[n] := \{1, 2, \ldots, n\}$ for every $n \in \mathbb{N}$. Furthermore, let $\operatorname{negl}(n)$ denote a negligible function, i.e., a function that is asymptotically smaller than the inverse of any polynomial. We follow the standard notation of quantum information [36]. We say that an algorithm $A$ is $\mathsf{QPT}$ if it is quantum and runs in polynomial time relative to its input size. In this work, when we say a polynomial, we always

mean a positive polynomial. We write $\mathrm{poly}(n)$ for an unspecified polynomial in $n$. A function $f$ is an *inverse polynomial* if $f(n) = 1/p(n)$ for some polynomial $p$. For $N \in \mathbb{N}$ and $q \in [0,1]$, we write $\mathrm{Bin}(N,q)$ for the binomial distribution with $N$ trials and success probability $q$. An algorithm $A$ is non-uniform if it is initialized with quantum advice and is said to be uniform if no advice is given. Adversaries in the security definitions and proofs are assumed to be uniform unless stated otherwise. For an algorithm $Q$ whose queries and responses are classical strings, we write $A^Q$ to mean that $A$ has classical query access to $Q$. When $Q = (Q_1, Q_2)$ is a pair of algorithms, $A^Q$ means that $A$ has classical query access to both $Q_1$ and $Q_2$; the two components may share an internal, possibly quantum, state, which each invocation may read and update. In all cases the interface is classical: we only consider classical oracle access in this work.

### 3.2 Indistinguishability Obfuscation for Null Quantum Circuits

We recall the notion of indistinguishability obfuscation for null quantum circuits from [13], restricting attention to quantum circuits producing a single classical output bit. Recall, a null quantum circuit is one that on any input, outputs 0 except with negligible probability.

Note that $\mathsf{qIO}$ has been studied and defined for more general quantum circuit classes [12, 10, 32]. We refer the reader to these works for the general definitions, but in this work we define $\mathsf{qIO}$ only for null quantum circuits.

**Definition 5.** *An indistinguishability obfuscator for null quantum circuits (null–qIO) consists of the following pair of* $\mathsf{QPT}$ *algorithms.*

- $\mathsf{Obf}(1^n, Q)$*: Given a security parameter* $1^n$ *and a classical description of a quantum circuit* $Q$ *with a single classical output bit, this algorithm produces a classical description of an obfuscated program* $\widetilde{Q}$*.*
- $\mathsf{Eval}(\widetilde{Q}, |\psi\rangle)$*: Given a classical description of a program* $\widetilde{Q}$ *and a quantum state* $|\psi\rangle$*, this algorithm produces a bit* $b$*.*

The correctness requirement is stated next.

**Definition 6 (Correctness).** *A null–qIO* $(\mathsf{Obf}, \mathsf{Eval})$ *satisfies correctness if, for every negligible function* $\nu$ *and every polynomial* $p$*, there exist a polynomial* $k$ *and a negligible function* $\mu$ *such that the following holds for every family* $\{Q_n\}_{n\in\mathbb{N}}$ *of quantum circuits with classical descriptions of size at most* $p(n)$*, states* $\{|\psi_n\rangle\}_{n\in\mathbb{N}}$*, and bits* $\{b_n\}_{n\in\mathbb{N}}$ *satisfying* $\Pr[Q_n(|\psi_n\rangle) = b_n] \geq 1 - \nu(n)$*:*

$$\Pr\Big[\mathsf{Eval}\big(\widetilde{Q}, |\psi_n\rangle^{\otimes k(n)}\big) = b_n : \widetilde{Q} \leftarrow \mathsf{Obf}(1^n, Q_n)\Big] \geq 1 - \mu(n). \tag{8}$$

*Remark 7.* This correctness notion relaxes the usual obfuscation guarantee along two dimensions. First, it only applies to inputs on which the circuit's behavior is decisive — accepting or rejecting with overwhelming probability. Second, it allows the evaluator to consume several copies of the input state rather than a single copy. Relaxing the definition strengthens our separation.

Finally, we state the security requirement.

**Definition 8 (Security of null–qIO).** *Let $\mathcal{O} = (\mathsf{Obf}, \mathsf{Eval})$ be a null–qIO scheme. We say that $\mathcal{O}$ is secure if the following holds.*

*Let $\{Q_{0,n}\}_{n\in\mathbb{N}}$ and $\{Q_{1,n}\}_{n\in\mathbb{N}}$ be any two polynomial-size families of quantum circuits having the same size and the same input and output spaces. Suppose that both families are null. Namely, suppose that there exists a negligible function $\nu$ such that, for every $b \in \{0,1\}$, every $n$, and every input state $\rho$, $\Pr[Q_{b,n}(\rho) = 1] \leq \nu(n)$.*

*Then, for any non-uniform $\mathsf{QPT}$ algorithm $D_n$,*

$$\left|\Pr[D_n(\mathsf{Obf}(1^n, Q_{0,n})) = 1] - \Pr[D_n(\mathsf{Obf}(1^n, Q_{1,n})) = 1]\right| \leq \mathrm{negl}(n).$$

### 3.3 Quantum Complexity Classes

We first recall the definitions of the complexity classes Quantum-Merlin-Arthur ($\mathsf{QMA}$) and Quantum-Classical-Interactive-Proofs ($\mathsf{QCIP}[2]$).

**Definition 9 ($\mathsf{QMA}$).** *A language $L \subseteq \{0,1\}^*$ is in $\mathsf{QMA}$ if there exists a $\mathsf{QPT}$ algorithm $Q$ and a polynomial $p$ such that for any $x \in \{0,1\}^*$,*

1. *If $x \in L$, then there exists a $p(|x|)$-qubit state $|w_x\rangle$ such that*

$$\Pr[1 \leftarrow Q(x, |w_x\rangle)] \;\geq\; \tfrac{2}{3}. \tag{9}$$

2. *If $x \notin L$, then for any $p(|x|)$-qubit state $|w\rangle$,*

$$\Pr[1 \leftarrow Q(x, |w\rangle)] \;\leq\; \tfrac{1}{3}. \tag{10}$$

Henceforth, for a $\mathsf{QMA}$ language $L$, we let $L_n := \{0,1\}^n \cap L$ and $\overline{L}_n := \{0,1\}^n \setminus L$.

**Definition 10 ($\mathsf{QMA}$ Relation).** *Let $L$ be a language in $\mathsf{QMA}$ and let $Q$ be its verification algorithm. A $\mathsf{QMA}$ relation $R_{L,Q}$ is the set of pairs $(x, |w\rangle)$ satisfying $\Pr[1 \leftarrow Q(x, |w\rangle)] \geq \frac{2}{3}$.*

We recall the complexity class $\mathsf{QCIP}$ as defined in [16, 29]. $\mathsf{QCIP}$ is the complexity class for interactive protocols over classical channels. We define it as a promise problem rather than a language, because the oracles of Definition 12 must be allowed to answer arbitrarily on ambiguous instances, that is, on instances whose protocol value lies strictly between the soundness and completeness thresholds. In this paper we focus only on $\mathsf{QCIP}[2]$, which is a two-message variant of $\mathsf{QCIP}$ and is defined as follows:

**Definition 11.** *A promise problem $(L_{\mathrm{yes}}, L_{\mathrm{no}})$ is in $\mathsf{QCIP}[2, c, s]$ if there exist $\mathsf{QPT}$ algorithms $V_1, V_2$ such that:*

– Completeness: *for every* $x \in L_{\text{yes}}$*, there is a family* $\{z_y : y \in \{0,1\}^{\text{poly}(|x|)}\}$ *of* $\text{poly}(|x|)$*-length strings with*

$$\Pr_{(\rho,y)\leftarrow V_1(x)}[V_2(x,\rho,y,z_y)=1] \;\geq\; c(|x|). \tag{11}$$

– Soundness: *for every* $x \in L_{\text{no}}$ *and every such family* $\{z_y\}$*,*

$$\Pr_{(\rho,y)\leftarrow V_1(x)}[V_2(x,\rho,y,z_y)=1] \;\leq\; s(|x|). \tag{12}$$

*Henceforth, we let* $\mathsf{QCIP}[2] := \mathsf{QCIP}[2, 2/3, 1/3]$ *to simplify notation.*

In the definition above, $\rho$ can be a quantum state, but $x, y, z_y$ are classical strings. Note that error reduction for $\mathsf{QCIP}[2]$ follows from parallel repetition: run independent copies of the protocol in parallel and accept by majority vote.

For $\mathsf{QPT}$ algorithms $V_1, V_2$ and a string $x$, we write

$$\text{val}_{V_1,V_2}(x) \;:=\; \max_{\{z_y\}_y} \Pr_{(\rho,y)\leftarrow V_1(x)}\big[V_2(x,\rho,y,z_y)=1\big] \tag{13}$$

for the value of the protocol on input $x$.

**Definition 12.** *The set* $\mathcal{O}_{\mathsf{QCIP}[2]}$ *consists of any (computationally unbounded) algorithm* $O$ *that satisfies the following condition. For any* $\mathsf{QPT}$ *algorithms,* $V_1$ *and* $V_2$*, and any* $x \in \{0,1\}^*$*,*

$$O(V_1,V_2,x) = \begin{cases} 1 & \text{if } \text{val}_{V_1,V_2}(x) \;\geq\; 2/3, \\ 0 & \text{if } \text{val}_{V_1,V_2}(x) \;\leq\; 1/3. \end{cases} \tag{14}$$

*On inputs with* $1/3 < \text{val}_{V_1,V_2}(x) < 2/3$ *the output of* $O$ *is unconstrained.*

Some care is needed when giving an algorithm with access to a $\mathsf{QCIP}[2]$ oracle, because $\mathsf{QCIP}[2]$ is a class of promise problems. We do not fix a single oracle but quantify over the whole set $\mathcal{O}_{\mathsf{QCIP}[2]}$. Concretely, throughout this work a statement of the form "$A^O$ cannot perform a task" (for instance, distinguishing two distributions) means: there exists at least one $O \in \mathcal{O}_{\mathsf{QCIP}[2]}$ for which $A^O$ fails. Symmetrically, $A^O$ counts as *succeeding* only if it succeeds against *every* $O \in \mathcal{O}_{\mathsf{QCIP}[2]}$. Condition 4 of Definition 25 is stated in exactly this form.

The reason for this convention is that on an ambiguous input $(V_1, V_2, x)$ whose protocol value lies strictly between the two thresholds, i.e. $1/3 < \text{val}_{V_1,V_2}(x) < 2/3$, it is not clear how the oracle should answer. This issue is not harmless: an oracle could smuggle useful information into those answers—information that has nothing to do with deciding the promise problem, yet that an algorithm might exploit. Quantifying over all valid $O$ rules this out. An algorithm that succeeds only under a particular choice of answers on ambiguous instances is reading off a feature of one implementation rather than using the power to decide promise instances, and does not count; only an algorithm that succeeds against every valid oracle is using that deciding power. This is why $A^O$ counts as *succeeding* only if it succeeds against *every* $O \in \mathcal{O}_{\mathsf{QCIP}[2]}$.

### 3.4 Falsifiable Assumptions

We recall the notion of falsifiable assumptions [38, 26, 3, 8].

**Definition 13 (Falsifiable Cryptographic Assumptions).** *A falsifiable cryptographic assumption is a pair $(C, c)$ of a* QPT *algorithm $C$ (challenger) and a constant $c \in [0, 1)$. $C$ interacts with an adversary over a quantum channel and then outputs $1/0$. The assumption is said to be true if, for any* QPT *algorithm $A$,*

$$\Pr\left[1 \leftarrow \langle C(1^n), A(1^n)\rangle\right] \ \leq \ c + \mathrm{negl}(n), \tag{15}$$

*where $1 \leftarrow \langle C(1^n), A(1^n)\rangle$ means that $C$ outputs $1$ after the interaction with $A$.*

### 3.5 Witness Encryption

We recall the notion of witness encryption (WE) for QMA from [13]. Initially, WE for NP was introduced in [24]. We define encryption only with respect to classical messages. Also, we only define it for single-bit messages, but it is easy to generalize the definition to multiple bits.

**Definition 14 (Witness Encryption for QMA).** *Let $L \in$ QMA and $Q$ be its verification algorithm. A WE scheme* $(\mathsf{WE.Enc}, \mathsf{WE.Dec})$ *for the relation $R_{L,Q}$ consists of the following* QPT *algorithms:*

- $\mathsf{WE.Enc}(1^n, x, m)$*: On input a statement $x \in \{0,1\}^n$ and a message $m \in \{0,1\}$, it outputs a classical ciphertext $ct$.*
- $\mathsf{WE.Dec}(x, ct, |\psi\rangle)$*: On input a statement $x$, a classical ciphertext $ct$, and a quantum state $|\psi\rangle$, it outputs a message $\widetilde{m} \in \{0,1\}$ or $\perp$.*

We define correctness below. Throughout, $t$ denotes the number of witness copies consumed by decryption.

**Definition 15 (Correctness).** *A WE* $(\mathsf{WE.Enc}, \mathsf{WE.Dec})$ *for a relation $R_{L,Q}$ is correct if there exists a negligible function $\nu(n)$ and a polynomial $t(n)$ such that for all $m \in \{0,1\}$, all sequences of instances $\{x_n\}_{n\in\mathbb{N}}$, where $|x_n| = n$ for each $n \in \mathbb{N}$, and witnesses $\{|\psi_n\rangle\}_{n\in\mathbb{N}}$, where each $(x_n, |\psi_n\rangle) \in R_{L,Q}$, it holds that*

$$\Pr\Big[\mathsf{WE.Dec}\big(x_n, \mathsf{WE.Enc}(1^n, x_n, m), |\psi_n\rangle^{\otimes t(n)}\big) = m\Big] \ \geq \ 1 - \nu(n).$$

Next we present the definition of security of WE against quantum algorithms. Our definition is a bit weaker than the traditional definition, but this just makes our separation stronger. Specifically, the traditional notion requires that for any $x \notin L$, encryptions of 0 and 1 are computationally indistinguishable. Instead, we allow the adversary to choose $x$, and it wins the security game if it manages to distinguish the encrypted message and $x \notin L$.

**Definition 16 (Security).** *A WE scheme $\Pi :=$* $(\mathsf{WE.Enc}, \mathsf{WE.Dec})$ *for a relation $R_{L,Q}$ is secure if for any* QPT *adversary $A := (A_1, A_2)$,*

$$\Pr[\mathsf{WE.Sec}_{\Pi,A}(1^n) = 1] \ \leq \ \tfrac{1}{2} + \mathrm{negl}(n). \tag{16}$$

$\mathsf{WE.Sec}_{\Pi,A}(1^n)$:

1. $b \leftarrow \{0,1\}$.
2. $x \leftarrow A_1(1^n)$.
3. $ct \leftarrow \mathsf{WE.Enc}(1^n, x, b)$.
4. $b' \leftarrow A_2(x, ct)$.
5. The output of the experiment is 1 if $x \in \overline{L}_n$ and $b' = b$. Otherwise, the output is 0.

**Theorem 17 (Lemma 3.15 in [13]).** *Assuming null–qIO, there exists WE for* QMA*.*

### 3.6 Quantum Black-Box Reductions

Quantum black-box reductions are explored in several works [35, 42, 20, 17]. We formulate quantum black-box reductions specifically showing the security of a WE scheme based on a falsifiable cryptographic assumption.

**Definition 18 ($\Pi$-adversary).** *Let $L \in$ QMA and $Q$ be its verification algorithm, and let $\Pi = (\mathsf{WE.Enc}, \mathsf{WE.Dec})$ be a WE scheme for the relation $R_{L,Q}$. We say that a quantum algorithm $\overline{P} = (\overline{P}_1, \overline{P}_2)$ is a $\Pi$-adversary if there exists a polynomial $p$ such that*

$$\Pr\left[\mathsf{WE.Sec}_{\Pi,\overline{P}}(1^n) = 1\right] \;\geq\; \tfrac{1}{2} + \tfrac{1}{p(n)} \tag{17}$$

*for infinitely many $n \in \mathbb{N}$. In this case, we say $\overline{P}$ has advantage $1/p$.*

**Definition 19 (WE–reduction).** *Let $\Pi$ be a WE scheme for* QMA*. A quantum black-box reduction showing the security of $\Pi$ from a falsifiable cryptographic assumption $(C, c)$ is a* QPT *algorithm $\Sigma^{(\cdot)}$ such that for any polynomial $p$, there exists a polynomial $p'$ such that for any (even inefficient) $\Pi$-adversary $\overline{P}$ with advantage $1/p$, the algorithm $\Sigma^{\overline{P}}$ breaks the assumption with advantage $1/p'$, i.e.*

$$\Pr\left[1 \leftarrow \langle C(1^n), \Sigma^{\overline{P}}(1^n)\rangle\right] \;\geq\; c + \frac{1}{p'(n)}, \tag{18}$$

*for infinitely many $n \in \mathbb{N}$. For short, we call such a reduction a* WE–reduction*.*

In this work, we consider only reductions $\Sigma$ that query the adversary on the same security parameter, non-adaptively, classically, and on honestly generated ciphertexts. More formally, our separation applies to reductions defined as follows, which we call *restricted WE–reduction*.

**Definition 20 (Restricted WE–reduction).** *Let $(C, c)$ be a falsifiable assumption, and let $\Sigma^{(\cdot)}$ be a WE–reduction defined in Definition 19. We say that $\Sigma^{(\cdot)}$ is a* restricted WE–reduction *if there exist polynomials $s = s(n)$ and $\ell = \ell(n)$ such that, in the interaction $\left\langle C(1^n), \Sigma^{\overline{P}}(1^n)\right\rangle$, the oracle access of $\Sigma$ to $\overline{P} = (\overline{P}_1, \overline{P}_2)$ has the following form.*

1. *The challenger $C$ and the reduction $\Sigma$ may interact arbitrarily throughout the execution. Only $\Sigma$ has oracle access to $\overline{P}$.*
2. *For each $i \in [s]$, $\Sigma$ submits $1^n$ to $\overline{P}_1$ and receives an instance $x_i$.*
3. *For each $i \in [s]$ and $j \in [\ell]$, a bit $m_{i,j} \leftarrow \{0,1\}$ is sampled uniformly at random and the ciphertext $\mathsf{ct}_{i,j} \leftarrow \mathsf{WE.Enc}(1^n, x_i, m_{i,j})$ is computed by running the encryption algorithm honestly. This step may be performed by $\Sigma$, by $C$, or jointly through their interaction.*
4. *For each $i \in [s]$ and $j \in [\ell]$, $\Sigma$ submits each $x_i$ and $ct_{i,j}$ to the oracle $\overline{P}_2$ and receives a response $\widetilde{m}_{i,j} \leftarrow \overline{P}_2(x_i, ct_{i,j})$.*
5. *The execution ends when $C$ outputs its decision bit.*

*Thus, the oracle queries are parameter-preserving, classical, and non-adaptive.*

*Remark 21 (Honestly generated ciphertexts).* Note that in Definition 20, we assume that ciphertext queries are honestly generated, meaning they are generated by running the encryption algorithm honestly on a randomly sampled plaintext bit. We believe that this restriction is justified because, intuitively, the reduction should turn a valid attack against the $\mathsf{WE}$ scheme into a valid attack against the falsifiable assumption. A valid attack against $\mathsf{WE}$ involves distinguishing or breaking the security of an honestly generated ciphertext. Therefore, a reduction should be able to use this property alone to carry out a successful attack on the underlying falsifiable assumption, without needing to query $\overline{P}$ on maliciously generated ciphertexts.

Since a restricted $\mathsf{WE}$–reduction only queries non-adaptively, we can split the execution in Definition 20 into multiple phases that will be useful in our analysis later on.

**Definition 22 (Decomposition of $\Gamma$).** *Let $(C, c)$ be a falsifiable assumption, and let $\Sigma^{(\cdot)}$ be a restricted $\mathsf{WE}$–reduction as in Definition 20. For any adversary $\overline{P} = (\overline{P}_1, \overline{P}_2)$, let*

$$\Gamma^{\overline{P}}(1^n) := \langle C(1^n), \Sigma^{\overline{P}}(1^n) \rangle \tag{19}$$

*denote the algorithm obtained by viewing $C$ and $\Sigma$ together as a single algorithm; its output is the decision bit of $C$. We define the following two algorithms, $\Gamma_1, \Gamma_2$, and consider the following* phased execution*:*

- ***Phase 1:*** *Run $\overline{P}_1(1^n)$ independently $s$ times to generate $x = (x_1, \ldots, x_s)$. The algorithm $\overline{P}_1$ is not executed in any later phase.*
- ***Phase 2:*** *$\Gamma_1(1^n, x)$ simulates the execution of $\Gamma^{\overline{P}}(1^n)$, including the interaction between $C$ and $\Sigma$, until just before $\Sigma$ makes its first query to $\overline{P}_2$. When $\Sigma$ makes its $i$-th query to $\overline{P}_1$, $\Gamma_1$ answers it with the pre-generated instance $x_i$ in Phase 1 instead of executing $\overline{P}_1$. The output of this phase consists of the pairs $(ct_{i,j})_{i \in [s], j \in [\ell]}$ generated in this execution, together with the joint internal state $st$ of $C$ and $\Sigma$ at this point.*
- ***Phase 3:*** *For each $i \in [s]$ and $j \in [\ell]$, run $\overline{P}_2$ on $(x_i, ct_{i,j})$, and let $\widetilde{m}_{i,j}$ be the response.*

– ***Phase 4:*** *$\Gamma_2(st, x, \mathsf{ct}, \widetilde{m})$ resumes the execution of $\Gamma^{\overline{P}}(1^n)$ from the state $st$. When $\Sigma$ queries $\overline{P}_2$ on $\mathsf{ct}_{i,j}$, $\Gamma_2$ answers it with the recorded response $\widetilde{m}_{i,j}$ obtained in Step 3 instead of executing $\overline{P}_2$. The phase ends when $C$ outputs its decision bit, which is the output of the phased execution.*

*Remark 23.* The state $st$ is determined by the execution of Phase 2. When no confusion arises, we suppress it from the notation and write $\Gamma_2(x, \mathsf{ct}, \widetilde{m})$.

**Lemma 24.** *Let $\Sigma^{(\cdot)}$ be a restricted $\mathsf{WE}$–reduction. Then the output of the phased execution of Definition 22 is distributed identically to the output of $\Gamma^{\overline{P}}(1^n)$.*

*Proof.* The phased execution differs from $\Gamma^{\overline{P}}(1^n)$ only in the timing of the executions of $\overline{P}_1$ and $\overline{P}_2$. Each execution of $\overline{P}_1$ takes the input $1^n$ and fresh randomness only, so its output is independent of the preceding transcript; the $s$ executions may therefore be moved to the beginning of the execution, and answering the queries of $\Sigma$ with the pre-generated $x_1, \ldots, x_s$ leaves the joint distribution of the execution unchanged. Similarly, each honest generation of $\mathsf{ct}_{i,j}$ is sampled from a distribution that depends only on $x_i$, so it may be moved before the first query to $\overline{P}_2$. Finally, since the queries to $\overline{P}_2$ are non-adaptive, the list $(x_i, \mathsf{ct}_{i,j})_{i,j}$ is determined at the end of Phase 2, and by assumption each response of $\overline{P}_2$ depends only on its query and on randomness independent of the transcript; hence computing all responses in Phase 3 and replaying them in Phase 4 also leaves the joint distribution unchanged. In particular, the decision bit of $C$ has the same distribution in both executions. □

### 3.7 QMA–QCIP[2] Gap Problem

We introduce the notion of QMA–QCIP[2] gap problems. We will assume the existence of these problems in our separation.

**Definition 25 (QMA–QCIP[2] gap problems).** *A QMA–QCIP[2] gap problem of a language $L$ in QMA with verification algorithm $Q$ and a QMA relation $R_{L,Q}$ is a pair $(\mathbf{L}, \overline{\mathbf{L}})$ satisfying the following conditions:*

1. *$\mathbf{L} := \{\mathbf{L}_n\}_{n\in\mathbb{N}}$ and $\overline{\mathbf{L}} := \{\overline{\mathbf{L}}_n\}_{n\in\mathbb{N}}$ are distribution ensembles, where $\mathbf{L}_n$ is a distribution over $L_n := L \cap \{0,1\}^n$ and $\overline{\mathbf{L}}_n$ is a distribution over $\overline{L}_n := \{0,1\}^n \setminus L_n$ for each $n \in \mathbb{N}$.*
2. *There exists a QPT algorithm SampYes that takes as input $(1^n, 1^t)$, for an arbitrary polynomial $t = t(n)$, and outputs $(x, |w_x\rangle^{\otimes t})$, where $(x, |w_x\rangle) \in R_{L,Q}$, such that the projection to the first coordinate of the output is $\mathbf{L}_n$.*
3. *There exists a not-necessarily-efficient algorithm SampNo that takes as input $1^n$ and samples from the distribution $\overline{\mathbf{L}}_n$.*
4. *For every polynomial $s$ and every QPT distinguisher $A$, there exists an oracle $O \in \mathcal{O}_{\mathsf{QCIP}[2]}$ such that*

$$\left| \Pr_{(x_1,\ldots,x_s)\leftarrow\mathbf{L}_n^{\otimes s}}[A^O(x_1,\ldots,x_s)=1] - \Pr_{(x_1,\ldots,x_s)\leftarrow\overline{\mathbf{L}}_n^{\otimes s}}[A^O(x_1,\ldots,x_s)=1] \right| \leq \mathrm{negl}(n), \tag{20}$$

*where $\mathbf{L}_n^{\otimes s}$ (resp. $\overline{\mathbf{L}}_n^{\otimes s}$) denotes $s$ independent samples of $\mathbf{L}_n$ (resp. $\overline{\mathbf{L}}_n$).*

*When $s$ independent instances are needed, we write* $\mathsf{SampYes}(1^n, 1^t, 1^s)$ *for the algorithm that runs $s$ independent copies of* $\mathsf{SampYes}(1^n, 1^t)$*, and similarly for* $\mathsf{SampNo}$.

Note that we only allow for classical access to the $\mathsf{QCIP}[2]$ oracle in this work. This can only strengthen our separation, as it weakens the assumption of a $\mathsf{QMA}$–$\mathsf{QCIP}[2]$ gap problem.

## 4 Simulating the Ideal Decryptor

In this section, we prove Theorem 26 below, which is the main simulation result used in our separation.

For any challenger $C$ of a falsifiable assumption and a restricted $\mathsf{WE}$–reduction $\Sigma$, let $\Gamma$ be the interaction between the two algorithms and let $\Gamma_1, \Gamma_2$ be the decomposition as defined in Definition 22.

Let $L$ be a $\mathsf{QMA}$ language with a $\mathsf{QMA}$–$\mathsf{QCIP}[2]$ gap problem $(\mathbf{L}, \overline{\mathbf{L}})$ and verification algorithm $Q$. Let $\Pi = (\mathsf{WE.Enc}, \mathsf{WE.Dec})$ be a $\mathsf{WE}$ scheme for $R_{L,Q}$ that uses at most $t = t(n)$ witness copies for decryption.

We now define the two experiments that will be compared. Both experiments depend on polynomials $s, \ell$ which are determined by $\Sigma$ in the same way as Definition 20. They also depend on $t$ which is determined by $\Pi$.

$\mathsf{DecYes}^{\mathsf{SampYes}}_{\Gamma,D}(1^n)$:

1. Sample
$$(x_i, |\psi_i\rangle^{\otimes t\ell})_{i\in[s]} \leftarrow \mathsf{SampYes}(1^n, 1^{t\ell}, 1^s),$$
and set $x = (x_1, \ldots, x_s)$.
2. Run
$$\left(ct_{i,j}\right)_{i,j} \leftarrow \Gamma_1(1^n, x).$$
3. For every $i \in [s]$ and $j \in [\ell]$, compute
$$a_{i,j} \leftarrow D(x_i, ct_{i,j}, |\psi_i\rangle^{\otimes t}),$$
using fresh witness copies for each query.
4. Output $\Gamma_2(x, ct, a)$.

**Fig. 1.** The yes-based experiment.

**Theorem 26.** *There exist a computationally unbounded classical function $S^*$ and a QPT algorithm $D^*$, both depending only on $\Pi$, such that the following holds.*

*Let $\epsilon$ be an inverse polynomial. For every falsifiable assumption $(C, c)$, and every restricted $\mathsf{WE}$–reduction $\Sigma$ showing the security of $\Pi$ from $(C, c)$, let $\Gamma$ be*

$\mathsf{DecNo}^{\mathsf{SampNo}}_{\Gamma,S}(1^n)$:

1. Sample
$$(x_1, \ldots, x_s) \leftarrow \mathsf{SampNo}(1^n, 1^s),$$
and set $x = (x_1, \ldots, x_s)$.
2. Run
$$\left(ct_{i,j}\right)_{i,j} \leftarrow \Gamma_1(1^n, x).$$
3. For every $i \in [s]$ and $j \in [\ell]$, compute
$$a_{i,j} \leftarrow S(x_i, ct_{i,j}).$$
4. Output $\Gamma_2(x, ct, a)$.

**Fig. 2.** The no-based experiment.

*the interaction $\langle C, \Sigma \rangle$ and let $(\Gamma^1, \Gamma^2)$ be its decomposition (see Definition 22). Then, for all sufficiently large $n$,*

$$\left| \Pr\left[\mathsf{DecYes}^{\mathsf{SampYes}}_{\Gamma,D^*}(1^n) = 1\right] - \Pr\left[\mathsf{DecNo}^{\mathsf{SampNo}}_{\Gamma,S^*}(1^n) = 1\right] \right| \leq \epsilon(n) \tag{21}$$

*for all sufficiently large $n \in \mathbb{N}$. In fact, $D^*$ can be taken as the decryption algorithm,* $\mathsf{WE.Dec}$.

Before proving Theorem 26, we establish the ingredients needed for the simulation.

### 4.1 The Bayes Decryptor

We now define $S^*$. It is determined entirely by $\mathsf{WE.Enc}$ and does not depend on the reduction, the challenger, or $\epsilon$.

**Definition 27 (Bayes decryptor and Bayes error).** *Let $x \in \{0,1\}^n$. Let $\boldsymbol{E}_x$ be the distribution of the pair $(m, ct)$ obtained by sampling a uniform bit $m \leftarrow \{0,1\}$ and then sampling $ct \leftarrow \mathsf{WE.Enc}(1^n, x, m)$.*

*For $b \in \{0,1\}$, define*

$$p_b(x, ct) := \frac{\Pr[\mathsf{WE.Enc}(1^n, x, b) = ct]}{\Pr[\mathsf{WE.Enc}(1^n, x, 0) = ct] + \Pr[\mathsf{WE.Enc}(1^n, x, 1) = ct]}. \tag{22}$$

*If the denominator is zero, set $p_0(x, ct) = p_1(x, ct) := 1/2$. Thus $p_b(x, ct)$ is the posterior probability that the plaintext was $b$, conditioned on observing $ct$. In particular,*

$$p_0(x, ct) + p_1(x, ct) = 1$$

*for every $ct$.*

*The* Bayes decryptor *is the deterministic classical function*

$$S^*(x, ct) := \arg\max_{b\in\{0,1\}} p_b(x, ct), \tag{23}$$

*where ties are broken towards* 0. *Its* Bayes error *is*

$$\Delta(x) := \Pr_{(m,ct)\leftarrow \boldsymbol{E}_x}\big[S^*(x, ct) \neq m\big]. \tag{24}$$

Conditioned on the ciphertext, $S^*$ errs with probability $\min_{b\in\{0,1\}} p_b(x, ct)$. Therefore,

$$\Delta(x) = \mathop{\mathbb{E}}_{ct}\left[\min_{b\in\{0,1\}} p_b(x, ct)\right] \leq \frac{1}{2}, \tag{25}$$

where $ct$ is distributed as the second coordinate of $\mathbf{E}_x$. Evaluating $S^*$ may require unbounded computation, which is allowed in Theorem 26.

The next two elementary lemmas will be used repeatedly.

**Lemma 28 (Optimality of $S^*$).** *Let $x \in \{0,1\}^n$, let $\tau_x$ be any quantum state depending only on $x$, and let $A$ be any, possibly computationally unbounded, quantum algorithm. Then*

$$\Pr_{(m,ct)\leftarrow \boldsymbol{E}_x}\big[A(x, ct, \tau_x) = m\big] \leq 1 - \Delta(x). \tag{26}$$

*Proof.* Condition on $x$ and $ct$. Since $\tau_x$ depends only on $x$, the plaintext $m$ is independent of $\tau_x$ given $(x, ct)$. Let $\alpha_b := \Pr[b \leftarrow A(x, ct, \tau_x)]$ for $b \in \{0, 1\}$. The conditional probability of recovering $m$ is therefore

$$\sum_{b\in\{0,1\}} \alpha_b p_b(x, ct) \leq \max_{b\in\{0,1\}} p_b(x, ct).$$

Averaging over $ct$ and using Eq. (25) proves the claim. □

**Lemma 29 (Per-message error).** *For every $x \in \{0,1\}^n$ and every $m \in \{0,1\}$,*

$$\Pr_{ct\leftarrow \mathsf{WE.Enc}(1^n,x,m)}\big[S^*(x, ct) \neq m\big] \leq 2\Delta(x). \tag{27}$$

*Proof.* By Eq. (24),

$$\Delta(x) = \frac{1}{2}\sum_{m\in\{0,1\}} \Pr_{ct\leftarrow \mathsf{WE.Enc}(1^n,x,m)}\big[S^*(x, ct) \neq m\big].$$

Both summands are non-negative, so each is at most $2\Delta(x)$. □

### 4.2 Decryptability on No-Instances

Our first main step is to show that the Bayes error is small for almost every no-instance drawn from $\overline{\mathbf{L}}_n$. Thus, although these instances have no witnesses, their honest ciphertexts can still be decrypted with small error by the possibly inefficient function $S^*$.

**Lemma 30 (Statistical decryptability on no-instances).** *For every inverse polynomial $\gamma$,*

$$\delta_{n,\gamma} := \Pr_{x \leftarrow \overline{\mathbf{L}}_n} \left[ \Delta(x) \geq \gamma(n) \right] \leq \mathrm{negl}(n). \tag{28}$$

*Proof.* We first show that the Bayes error is negligible on every yes-instance.

**Step 1: $\Delta(x) \leq \nu(n)$ for every $x \in L \cap \{0,1\}^n$.** Let $x \in L \cap \{0,1\}^n$, and let $|w_x\rangle$ satisfy $(x, |w_x\rangle) \in R_{L,Q}$. Apply Lemma 28 with $\tau_x = |w_x\rangle^{\otimes t}$ and $A = \mathsf{WE.Dec}$. By correctness, $\mathsf{WE.Dec}$ recovers a uniformly random plaintext with probability at least $1 - \nu(n)$. Hence

$$1 - \nu(n) \leq 1 - \Delta(x),$$

and therefore $\Delta(x) \leq \nu(n)$.

**Step 2: a two-message protocol.** Fix $\gamma$ and set

$$r := \left\lceil \frac{16}{\gamma(n)} \right\rceil.$$

Since $\gamma$ is inverse polynomial, $r$ is polynomial. Define $\mathsf{QPT}$ algorithms $V_1, V_2$ as follows:

- $V_1(x)$: for each $j \in [r]$, sample $m_j \leftarrow \{0,1\}$ and $ct_j \leftarrow \mathsf{WE.Enc}(1^n, x, m_j)$. Send
$$y = (ct_1, \ldots, ct_r)$$
and keep
$$\rho = (m_1, \ldots, m_r)$$
private.
- $V_2(x, \rho, y, z_y)$: parse $z_y = (b_1, \ldots, b_r) \in \{0,1\}^r$ and accept if and only if
$$\left| \{ j \in [r] : b_j \neq m_j \} \right| \leq \frac{\gamma r}{2}.$$

**Step 3: the value of the protocol.** We claim that, for every $x$,

$$\mathrm{val}_{V_1,V_2}(x) = \Pr\left[ Z \leq \frac{\gamma r}{2} \right], \qquad Z \sim \mathrm{Bin}(r, \Delta(x)). \tag{29}$$

Fix a verifier message $y = (ct_1, \ldots, ct_r)$ and write

$$\mu_j := \min_{b \in \{0,1\}} p_b(x, ct_j).$$

The pairs $(m_j, ct_j)$ are independent samples from $\mathbf{E}_x$. Therefore, conditioned on $y$, the plaintext bits $m_1, \ldots, m_r$ remain independent and

$$\Pr[m_j = b \mid y] = p_b(x, ct_j).$$

Now fix a response $z = (b_1, \ldots, b_r)$. Conditioned on $y$, the error in position $j$ has probability

$$q_j := 1 - p_{b_j}(x, ct_j) \geq \mu_j,$$

with equality when $b_j = S^*(x, ct_j)$. Let $U_1, \ldots, U_r$ be independent uniform random variables on $[0, 1]$ and set

$$X_j = \mathbf{1}[U_j \leq q_j], \qquad X_j^* = \mathbf{1}[U_j \leq \mu_j].$$

Then $X_j^* \leq X_j$ pointwise for every $j$. Hence

$$\Pr\left[\sum_j X_j \leq \frac{\gamma r}{2} \,\middle|\, y\right] \leq \Pr\left[\sum_j X_j^* \leq \frac{\gamma r}{2} \,\middle|\, y\right].$$

Thus, for every fixed $y$, the optimal response is

$$z_y^* = \big(S^*(x, ct_1), \ldots, S^*(x, ct_r)\big).$$

By the definition of $\mathsf{QCIP}[2]$, the prover may choose its response separately for each verifier message $y$, so this pointwise choice attains the protocol value.

Under this strategy, the error indicators are $\mathbf{1}[S^*(x, ct_j) \neq m_j]$. Before conditioning on $y$, these are independent Bernoulli variables with mean $\Delta(x)$. Their sum is therefore distributed as $\mathrm{Bin}(r, \Delta(x))$, proving Eq. (29).

**Step 4: the two cases.** For all sufficiently large $n$:

- If $x \in L \cap \{0,1\}^n$, then $\Delta(x) \leq \nu(n)$ by Step 1. Hence $\mathbb{E}[Z] \leq \nu r$, and Markov's inequality gives

$$\Pr\left[Z > \frac{\gamma r}{2}\right] \leq \frac{2\nu}{\gamma} = \mathrm{negl}(n).$$

  Therefore $\mathrm{val}_{V_1,V_2}(x) \geq 2/3$.
- If $x \notin L$ and $\Delta(x) \geq \gamma$, then $\mathbb{E}[Z] = \Delta(x)r \geq \gamma r$. Thus $\gamma r/2 \leq \mathbb{E}[Z]/2$, and the multiplicative Chernoff bound gives

$$\Pr\left[Z \leq \frac{\gamma r}{2}\right] \leq e^{-\mathbb{E}[Z]/8} \leq e^{-\gamma r/8} \leq e^{-2} < \frac{1}{3}.$$

  Hence $\mathrm{val}_{V_1,V_2}(x) \leq 1/3$.

**Step 5: applying the gap assumption.** Let $A^O(x)$ submit $(V_1, V_2, x)$ to the oracle $O$ and output the returned bit. This is a $\mathsf{QPT}$ algorithm making one

classical oracle query. By Definition 12 and Step 4, for all sufficiently large $n$ and every $O \in \mathcal{O}_{\mathsf{QCIP}[2]}$,

$$\Pr_{x \leftarrow \mathbf{L}_n}[A^O(x) = 1] = 1$$

and

$$\Pr_{x \leftarrow \overline{\mathbf{L}}_n}[A^O(x) = 1] \leq 1 - \delta_{n,\gamma}.$$

Condition 4 of Definition 25 gives an oracle $O \in \mathcal{O}_{\mathsf{QCIP}[2]}$ for which these two probabilities differ by at most a negligible amount. For that oracle their difference is at least $\delta_{n,\gamma}$. Therefore

$$\delta_{n,\gamma} \leq \mathrm{negl}(n),$$

as required. □

### 4.3 Estimating the Ideal Experiment

Fix the two-stage process $\Gamma = (\Gamma_1, \Gamma_2)$ defined at the beginning of this section. Recall that every real plaintext is sampled as a fresh uniform bit. More precisely, for every $i \in [s]$ and $j \in [\ell]$, the corresponding real ciphertext is generated as

$$\widetilde{m}_{i,j} \leftarrow \{0,1\}, \qquad ct_{i,j} \leftarrow \mathsf{WE.Enc}(1^n, x_i, \widetilde{m}_{i,j}).$$

For an instance tuple $x = (x_1, \ldots, x_s)$, define the *ideal experiment* $\mathsf{Ideal}_\Gamma(x)$ as follows. Run

$$(ct_{i,j})_{i,j} \leftarrow \Gamma_1(1^n, x),$$

and let $\widetilde{m} = (\widetilde{m}_{i,j})_{i,j}$ be the uniform plaintext bits used to generate the real ciphertexts in this execution. Instead of querying a decryptor, give $\widetilde{m}$ to the second stage in its original $(i,j)$ positions and output

$$\Gamma_2(x, ct, \widetilde{m}).$$

Write

$$\iota(x) := \Pr[\mathsf{Ideal}_\Gamma(x) = 1]. \tag{30}$$

The ideal experiment is only a proof device. In particular, the plaintext vector need not be available to the reduction or to an efficient algorithm. We next show that $\iota(x)$ can nevertheless be estimated by a $\mathsf{QPT}$ algorithm with access to a $\mathsf{QCIP}[2]$ oracle.

To do this, the verifier hides the real ciphertexts among test ciphertexts. It does not need to know the real plaintexts. In each block, it adds the same number of test encryptions of 0 and of 1.

Fix an even polynomial $K = K(n)$ with $K \geq \ell$. Define $\mathsf{QPT}$ algorithms $(V_1^\Gamma, V_2^\Gamma)$ as follows.

$V_1^{\Gamma}(x)$ on input $x = (x_1, \ldots, x_s)$:

1. Run
$$(ct_{i,j})_{i,j} \leftarrow \Gamma_1(1^n, x).$$
These are the *real* ciphertexts. Their plaintexts are not needed by the verifier.
2. For every $i \in [s]$, generate $K/2$ test encryptions of 0 and $K/2$ test encryptions of 1:
$$ct'_{i,c,r} \leftarrow \mathsf{WE.Enc}(1^n, x_i, c), \qquad c \in \{0,1\}, \quad r \in [K/2].$$
Every test ciphertext is generated honestly.
3. For every $i \in [s]$, choose an independent uniformly random permutation $\pi_i$ of the $K + \ell$ ciphertexts in block $i$. Let $y_i$ be the resulting shuffled list.
4. Send $y = (y_1, \ldots, y_s)$. Keep the permutations, the real and test positions, and the plaintexts of the test ciphertexts private in a state $\rho$.

$V_2^{\Gamma}(x, \rho, y, z_y)$: parse $z_y$ as one bit for every shuffled ciphertext. Reject if any test ciphertext is answered incorrectly, i.e. does not match its plaintext. Otherwise, extract the answers corresponding to the real ciphertexts, restore their original $(i, j)$ order, call the resulting vector $a$, and output $\Gamma_2(x, ct, a)$.

**Lemma 31 (Value of the hidden-test protocol).** *For every $x = (x_1, \ldots, x_s)$,*

$$\mathrm{val}_{V_1^{\Gamma}, V_2^{\Gamma}}(x) \le \iota(x) + \frac{2s\ell}{K}. \tag{31}$$

*Moreover, if $\Delta(x_i) \le \gamma$ for every $i \in [s]$, then*

$$\mathrm{val}_{V_1^{\Gamma}, V_2^{\Gamma}}(x) \ge \iota(x) - s(K + \ell)\gamma \ge \iota(x) - 2sK\gamma. \tag{32}$$

*Proof.* Let $W$ be the event that every answer, on both real and test ciphertexts, equals the corresponding true plaintext.

Couple the protocol and the ideal experiment so that they use the same execution of $\Gamma_1$. In particular, they use the same real ciphertexts, the same latent plaintext vector $\widetilde{m}$. On the event $W$, the vector supplied to $\Gamma_2$ is exactly $\widetilde{m}$ in its original order. Therefore,

$$\Pr[V_2^{\Gamma} \text{ accepts} \wedge W] = \Pr[\mathsf{Ideal}_{\Gamma}(x) = 1 \wedge W]. \tag{33}$$

*Lower bound.* Consider the prover that answers every ciphertext $c$ in block $i$ with $S^*(x_i, c)$. For $c \in \{0, 1\}$, write

$$e_{i,c} := \Pr_{ct \leftarrow \mathsf{WE.Enc}(1^n, x_i, c)}[S^*(x_i, ct) \neq c].$$

By the definition of the Bayes error,

$$\Delta(x_i) = \frac{e_{i,0} + e_{i,1}}{2}.$$

At each real position in block $i$, conditioned on the preceding transcript, the plaintext is uniform and the ciphertext is a fresh honest encryption. Consequently, the error probability at that position is $\Delta(x_i)$, and the sum of the error probabilities at the $\ell$ real positions is $\ell\Delta(x_i)$. The block also contains $K/2$ tests of each bit, whose error probabilities sum to $K\Delta(x_i)$. Thus, a union bound over all ciphertexts gives

$$\Pr[\neg W] \le \sum_{i=1}^{s} (K+\ell)\Delta(x_i) \le s(K+\ell)\gamma \le 2sK\gamma,$$

where the last inequality uses $K \ge \ell$. Using Eq. (33),

$$\begin{aligned} \mathrm{val}_{V_1^\Gamma, V_2^\Gamma}(x) &\ge \Pr[V_2^\Gamma \text{ accepts} \wedge W] && (34)\\ &= \Pr[\mathsf{Ideal}_\Gamma(x) = 1 \wedge W] && (35)\\ &\ge \iota(x) - s(K+\ell)\gamma. && (36) \end{aligned}$$

*Upper bound.* Fix an arbitrary prover strategy. Since the verifier rejects whenever a test ciphertext is answered incorrectly,

$$\begin{aligned} \Pr[\text{accept}] &\le \Pr[\text{accept} \wedge W] + \Pr[\text{all tests correct} \wedge \neg W] && (37)\\ &\le \iota(x) + \Pr[\mathcal{B}], && (38) \end{aligned}$$

where $\mathcal{B}$ is the event that every test ciphertext is answered correctly but at least one real ciphertext is answered incorrectly.

Fix a block $i$. For $c \in \{0,1\}$, let

$$R_{i,c} := \big|\{j \in [\ell] : \widetilde{m}_{i,j} = c\}\big|.$$

This quantity is used only in the analysis. The verifier does not need to know it.

For the purpose of an upper bound, reveal to the prover the plaintext label of every ciphertext in every block. This can only make the prover stronger. Since its response is classical, randomizing the response cannot improve its success probability. Condition on the public message $y$, all these plaintext labels, and a fixed response. This fixes its error positions and the values $R_{i,c}$.

There are $K/2 + R_{i,c}$ ciphertexts with plaintext $c$ in block $i$. Exactly $R_{i,c}$ of them are real. Conditioned on the ciphertexts and their plaintext labels, the real positions form a uniformly random $R_{i,c}$-subset of these $K/2 + R_{i,c}$ positions. Indeed, all real and test ciphertexts with plaintext $c$ are independent samples from the same distribution $\mathsf{WE.Enc}(1^n, x_i, c)$, and the final permutation is uniform.

If the prover makes at least one error in block $i$, choose one erroneous position $u$ according to a fixed rule, and let $c$ be its plaintext. If all tests in block $i$ are

nevertheless answered correctly, then $u$ must be a real position. Hence,

$$\Pr[\text{some real answer in block } i \text{ is wrong } \wedge \text{ all tests in block } i \text{ are correct} \mid y,\ \text{labels}, z_y] \tag{39}$$

$$\leq \Pr[u \text{ is real} \mid y,\ \text{labels}, z_y] \tag{40}$$

$$= \frac{R_{i,c}}{K/2 + R_{i,c}} \leq \frac{\ell}{K/2 + \ell} \leq \frac{2\ell}{K}. \tag{41}$$

If the prover makes no error in block $i$, the event on the left is empty and the same bound holds.

Since $z_y$ depends only on the revealed view and not on the hidden origins, conditioning on it does not change the uniform-subset property. Averaging over $y$, the plaintext labels, and the response preserves the bound. Finally, $\mathcal{B}$ implies that this event occurs in at least one block. A union bound over $i \in [s]$ gives

$$\Pr[\mathcal{B}] \leq \frac{2s\ell}{K}.$$

Combining this with Eq. (38) proves Eq. (31). □

To recover the numerical value of the protocol using only the decision oracle, we query several thresholds. Fix a positive inverse polynomial $\eta$. For all sufficiently large $n$, assume $\eta \leq 1$. Set

$$M := \left\lceil \frac{1}{\eta^2} \right\rceil.$$

For $\theta \in [0,1]$, let $(V_1^{\Gamma,\theta}, V_2^{\Gamma,\theta})$ be the protocol that runs $M$ independent copies of $(V_1^{\Gamma}, V_2^{\Gamma})$ in parallel and accepts if at least $\theta M$ of them accept.

**Lemma 32 (Value of the repeated protocol).** *For every $x$ and every $\theta \in [0,1]$,*

$$\mathrm{val}_{V_1^{\Gamma,\theta}, V_2^{\Gamma,\theta}}(x) = \Pr[\mathrm{Bin}(M, v) \geq \theta M], \qquad v := \mathrm{val}_{V_1^{\Gamma}, V_2^{\Gamma}}(x). \tag{42}$$

*Proof.* For a verifier message $y$, define

$$q(y) := \max_{z_y} \Pr[V_2^{\Gamma}(x, \rho, y, z_y) = 1 \mid y].$$

Then

$$v = \mathbb{E}_y[q(y)].$$

Now consider the $M$-fold protocol and fix all verifier messages $y^{(1)}, \ldots, y^{(M)}$. Conditioned on these messages, the private verifier states are independent. For any fixed response vector, the $M$ acceptance indicators are therefore independent. Their success probabilities depend coordinatewise on the responses supplied to the corresponding copies.

The prover's response is classical, and its success probability is linear in any distribution over response vectors. Thus, it suffices to consider a fixed response vector for every tuple of verifier messages.

The probability that at least $\theta M$ copies accept is nondecreasing in each of these success probabilities. Hence an optimal prover maximizes every coordinate separately. It attains success probability $q(y^{(r)})$ in copy $r$. A response in one copy cannot improve the acceptance probability of a different copy.

After averaging over the independent verifier messages, the $M$ acceptance indicators are independent and identically distributed Bernoulli variables. Each has success probability

$$\mathbb{E}_y[q(y)] = v.$$

Their sum therefore has distribution $\mathrm{Bin}(M, v)$, proving the claim. □

We can now define the estimator. Let

$$J := \left\lfloor \frac{1}{\eta} \right\rfloor.$$

$\widetilde{\Gamma}^O(x)$:

1. For every $j \in \{0, 1, \ldots, J\}$, submit
$$\big(V_1^{\Gamma, j\eta}, V_2^{\Gamma, j\eta}, x\big)$$
to the oracle $O$, and let
$$c_j := O\big(V_1^{\Gamma, j\eta}, V_2^{\Gamma, j\eta}, x\big)$$
be its output.
2. Set
$$\widehat{v} := \eta \max\{j : c_j = 1\}.$$
3. Output 1 with probability $\widehat{v}$, and output 0 otherwise.

**Lemma 33 (Accuracy of the estimator).** *For every $O \in \mathcal{O}_{\mathsf{QCIP}[2]}$ and every $x$,*

$$\left| \Pr[\widetilde{\Gamma}^O(x) = 1] - \mathrm{val}_{V_1^\Gamma, V_2^\Gamma}(x) \right| \le 2\eta. \tag{43}$$

*Moreover, $\widetilde{\Gamma}$ is $\mathsf{QPT}$ and makes $J + 1$ classical queries to $O$.*

*Proof.* Write

$$v := \mathrm{val}_{V_1^\Gamma, V_2^\Gamma}(x).$$

By Lemma 32 and Hoeffding's inequality, if $v \ge \theta + \eta$, then

$$\mathrm{val}_{V_1^{\Gamma,\theta}, V_2^{\Gamma,\theta}}(x) \ge 1 - e^{-2\eta^2 M} \ge 1 - e^{-2} > \frac{2}{3}.$$

Similarly, if $v \le \theta - \eta$, then

$$\mathrm{val}_{V_1^{\Gamma,\theta}, V_2^{\Gamma,\theta}}(x) \le e^{-2\eta^2 M} \le e^{-2} < \frac{1}{3}.$$

Therefore, by Definition 12, every valid oracle $O$ satisfies

$$c_j = 1 \quad \text{whenever} \quad j\eta \le v - \eta,$$

and

$$c_j = 0 \quad \text{whenever} \quad j\eta \ge v + \eta.$$

We do not use the oracle's answers at the remaining thresholds.

The query with $j = 0$ always returns $c_0 = 1$, since the repeated protocol with threshold 0 accepts unconditionally. Hence $\widehat{v}$ is well defined. Every $j$ with $c_j = 1$ satisfies $j\eta < v + \eta$, so

$$\widehat{v} \le v + \eta.$$

If $v \ge \eta$, let $j^\star$ be the largest index satisfying $j^\star\eta \le v - \eta$. Such an index belongs to $\{0, \dots, J\}$. By maximality,

$$c_{j^\star} = 1 \qquad \text{and} \qquad j^\star\eta > v - 2\eta.$$

If $v < \eta$, then $\widehat{v} \ge 0 > v - 2\eta$. Thus, in all cases,

$$|\widehat{v} - v| \le 2\eta.$$

Since the estimator outputs 1 with probability $\widehat{v}$, this proves the first claim.

Finally, $J$ and $M$ are polynomial because $\eta$ is inverse polynomial. Each queried verifier is $\mathsf{QPT}$, so $\widetilde{\Gamma}$ is $\mathsf{QPT}$ and makes exactly $J + 1$ classical oracle queries. □

### 4.4 Proof of Theorem 26

*Proof.* Fix the induced two-stage process $\Gamma = (\Gamma_1, \Gamma_2)$ and set

$$K := 2\left\lceil \frac{256 s^2 \ell}{\epsilon} \right\rceil, \qquad \gamma := \frac{\epsilon}{512 s^2 K}, \qquad \eta := \frac{\epsilon}{256 s}. \tag{44}$$

For all sufficiently large $n$, $K \ge \ell$. The parameters $K$, $M = \lceil 1/\eta^2 \rceil$, and $J = \lfloor 1/\eta \rfloor$ are polynomial, while $\gamma$ and $\eta$ are inverse polynomial.

Call an instance $x$ *good* if $\Delta(x) \le \gamma(n)$. By Step 1 of the proof of Lemma 30, every $x \in L \cap \{0,1\}^n$ is good for all sufficiently large $n$, because $\nu(n) \le \gamma(n)$ eventually. By Lemma 30,

$$\Pr_{x \leftarrow \overline{\mathbf{L}}_n}[x \text{ is not good}] \le \mathrm{negl}(n).$$

Hence a tuple sampled from either $\mathbf{L}_n^{\otimes s}$ or $\overline{\mathbf{L}}_n^{\otimes s}$ has all coordinates good except with negligible probability.

We use the following coupling observation repeatedly. The yes, no, and ideal experiments all use the same first-stage process $\Gamma_1$ and then apply the same second-stage process $\Gamma_2$ to an answer vector. Whenever two experiments produce the same answer vector, the input to $\Gamma_2$ is identical. Therefore the difference between their output probabilities is at most the probability that their answer vectors differ.

**Claim 1.**

$$\left|\Pr[\mathsf{DecYes}^{\mathsf{SampYes}}_{\Gamma,D}(1^n)=1]-\mathop{\mathbb{E}}_{x\leftarrow\mathbf{L}_n^{\otimes s}}[\iota(x)]\right|\le s\ell\,\nu(n). \tag{45}$$

In the yes experiment, the answer to query $(i,j)$ is

$$a_{i,j}=D^*(x_i,ct_{i,j},|\psi_i\rangle^{\otimes t}),$$

while the corresponding ideal answer is $\widetilde{m}_{i,j}$. At the moment $ct_{i,j}$ is generated, conditioned on the preceding transcript and on $\widetilde{m}_{i,j}$, it is an honest encryption under $x_i$. Moreover, $\Gamma_1$ never receives the witness copies. Hence correctness gives

$$\Pr[a_{i,j}\neq\widetilde{m}_{i,j}]\le\nu(n),$$

even after conditioning on the preceding transcript. A union bound over the $s\ell$ queries and the coupling observation proves Eq. (45).

**Claim 2.**

$$\left|\Pr[\mathsf{DecNo}^{\mathsf{SampNo}}_{\Gamma,S^*}(1^n)=1]-\mathop{\mathbb{E}}_{x\leftarrow\overline{\mathbf{L}}_n^{\otimes s}}[\iota(x)]\right|\le 2s\ell\gamma+\mathrm{negl}(n). \tag{46}$$

By Lemma 30 and a union bound over $i\in[s]$, all coordinates of $x\leftarrow\overline{\mathbf{L}}_n^{\otimes s}$ are good except with negligible probability. Condition on an all-good tuple. At the moment $ct_{i,j}$ is generated, conditioned on the preceding transcript and its plaintext $\widetilde{m}_{i,j}$, it is an honest encryption under $x_i$. Therefore, by Lemma 29,

$$\Pr[S^*(x_i,ct_{i,j})\neq\widetilde{m}_{i,j}\mid\text{preceding transcript}]\le 2\Delta(x_i)\le 2\gamma.$$

A union bound over the $s\ell$ queries and the coupling observation proves Eq. (46).

**Claim 3.**

$$\left|\mathop{\mathbb{E}}_{x\leftarrow\mathbf{L}_n^{\otimes s}}[\iota(x)]-\mathop{\mathbb{E}}_{x\leftarrow\overline{\mathbf{L}}_n^{\otimes s}}[\iota(x)]\right|\le\frac{\epsilon}{16}+\mathrm{negl}(n). \tag{47}$$

Let $\widetilde{\Gamma}$ be the estimator defined above. On every all-good tuple, Lemmas 31 and 33 imply, for every $O\in\mathcal{O}_{\mathsf{QCIP}[2]}$,

$$\left|\Pr[\widetilde{\Gamma}^O(x)=1]-\iota(x)\right|\le 2\eta+4sK\gamma+\frac{s\ell}{K}=:\beta. \tag{48}$$

By Eq. (44),

$$2\eta = \frac{\epsilon}{128s}, \qquad 4sK\gamma = \frac{\epsilon}{128s},$$

and

$$\frac{s\ell}{K} \leq \frac{\epsilon}{256s}.$$

Therefore

$$\beta \leq \frac{5\epsilon}{256s} \leq \frac{\epsilon}{32s}.$$

Apply condition 4 of Definition 25 to the $\mathsf{QPT}$ distinguisher $\widetilde{\Gamma}$. There exists an oracle $O \in \mathcal{O}_{\mathsf{QCIP}[2]}$ such that

$$\left| \Pr_{x \leftarrow \mathbf{L}_n^{\otimes s}}[\widetilde{\Gamma}^O(x) = 1] - \Pr_{x \leftarrow \overline{\mathbf{L}}_n^{\otimes s}}[\widetilde{\Gamma}^O(x) = 1] \right| \leq \mathrm{negl}(n). \tag{49}$$

For this same oracle, Eq. (48) holds on every all-good tuple. All-good tuples have all but negligible probability under both distributions. Therefore,

$$\left| \Pr_{x \leftarrow \mathbf{L}_n^{\otimes s}}[\widetilde{\Gamma}^O(x) = 1] - \mathop{\mathbb{E}}_{x \leftarrow \mathbf{L}_n^{\otimes s}}[\iota(x)] \right| \leq \beta + \mathrm{negl}(n),$$

and the same bound holds with $\mathbf{L}_n^{\otimes s}$ replaced by $\overline{\mathbf{L}}_n^{\otimes s}$. Combining these two bounds with Eq. (49) gives

$$\left| \mathop{\mathbb{E}}_{x \leftarrow \mathbf{L}_n^{\otimes s}}[\iota(x)] - \mathop{\mathbb{E}}_{x \leftarrow \overline{\mathbf{L}}_n^{\otimes s}}[\iota(x)] \right| \leq 2\beta + \mathrm{negl}(n) \leq \frac{\epsilon}{16} + \mathrm{negl}(n),$$

which proves Eq. (47).

Combining Eqs. (45) to (47) by the triangle inequality, and using

$$2s\ell\gamma = \frac{\epsilon\ell}{256sK} \leq \frac{\epsilon}{256},$$

we obtain

$$\left| \Pr[\mathsf{DecYes}^{\mathsf{SampYes}}_{\Gamma,D^*}(1^n) = 1] - \Pr[\mathsf{DecNo}^{\mathsf{SampNo}}_{\Gamma,S^*}(1^n) = 1] \right| \tag{50}$$

$$\leq s\ell\,\nu(n) + \frac{\epsilon}{256} + \frac{\epsilon}{16} + \mathrm{negl}(n). \tag{51}$$

Since $\epsilon$ is inverse polynomial while $s\ell\,\nu(n)$ and the remaining error are negligible, the right-hand side is at most $\epsilon(n)$ for all sufficiently large $n$. This proves Theorem 26. □

## 5 Separation

We now use the results of Section 4 to show that oracle access to the inefficient $\mathsf{WE}$ adversary can be simulated by an efficient algorithm.

**Theorem 34.** *Let $\epsilon = \epsilon(n)$ be any positive inverse polynomial. Let $L$ be a $\mathsf{QMA}$ language with verification algorithm $Q$ and a $\mathsf{QMA}$–$\mathsf{QCIP}[2]$ gap problem, and let $\Pi = (\mathsf{WE.Enc}, \mathsf{WE.Dec})$ be a $\mathsf{WE}$ scheme for $R_{L,Q}$. Then there exists an algorithm $\overline{P}$ satisfying the following conditions:*

- *$\overline{P}$ is a $\Pi$-adversary with advantage $1/4$.*
- *Oracle access to $\overline{P}$ is efficiently simulatable. In particular, for every falsifiable assumption $(C, c)$ and every restricted $\mathsf{WE}$–reduction $\Sigma$ showing the security of $\Pi$ from $(C, c)$, there exists a QPT algorithm $P$ such that*

$$\left| \Pr[\Gamma^{\overline{P}}(1^n) = 1] - \Pr[\Gamma^{P}(1^n) = 1] \right| \leq \epsilon(n) \tag{52}$$

*for all sufficiently large $n$, where $\Gamma^{\overline{P}}(1^n) := \langle C(1^n), \Sigma^{\overline{P}}(1^n) \rangle$ and $\Gamma^{P}(1^n) := \langle C(1^n), \Sigma^{P}(1^n) \rangle$.*

*Proof.* Let $\overline{P} = (\overline{P}_1, \overline{P}_2)$ be the following possibly inefficient algorithm. On input $1^n$, $\overline{P}_1$ samples

$$x \leftarrow \mathsf{SampNo}(1^n)$$

and outputs $x$. On input $(x, ct)$, $\overline{P}_2$ outputs

$$m \leftarrow S^*(x, ct),$$

where $S^*$ is the classical function of Definition 27. Note that $\overline{P}$ depends neither on $\Sigma$ nor on $\epsilon$.

Now fix any restricted $\mathsf{WE}$–reduction $\Sigma$. Let $P = (P_1, P_2)$ be the following $\mathsf{QPT}$ algorithm. On each invocation with input $1^n$, $P_1$ samples

$$(x, |w_x\rangle^{\otimes t\ell}) \leftarrow \mathsf{SampYes}(1^n, 1^{t\ell}),$$

outputs $x$, and keeps the witness copies as its private state, where $t$ is the number of witness copies used in decryption.

On input $(x, ct)$, $P_2$ computes

$$m \leftarrow \mathsf{WE.Dec}(x, ct, |w_x\rangle^{\otimes t}),$$

consuming $t$ witness copies for each query. Since a restricted $\mathsf{WE}$–reduction makes at most $\ell$ ciphertext queries for each instance, the $t\ell$ witness copies are sufficient.

**$\overline{P}$ is a $\Pi$-adversary.** Consider the security game $\mathsf{WE.Sec}_{\Pi,\overline{P}}(1^n)$. A bit $b \leftarrow \{0,1\}$ is sampled, $\overline{P}_1$ samples an instance $x \leftarrow \mathsf{SampNo}(1^n)$, and the challenger computes

$$ct \leftarrow \mathsf{WE.Enc}(1^n, x, b).$$

Then $\overline{P}_2$ outputs $b' \leftarrow S^*(x, ct)$.

The game outputs 1 exactly when $b' = b$, since $\mathsf{SampNo}$ outputs $x \in \overline{L}_n$ with probability 1. By the definition of $S^*$, this happens with probability $1 - \Delta(x)$. Therefore,

$$\Pr\left[\mathsf{WE.Sec}_{\Pi,\overline{P}}(1^n) = 1\right] = 1 - \mathop{\mathbb{E}}_{x \leftarrow \mathsf{SampNo}(1^n)}\left[\Delta(x)\right]. \tag{53}$$

Apply Lemma 30 with $\gamma(n) := 1/n$. Since $\Delta(x) \leq 1/2$ by Eq. (25),

$$\mathop{\mathbb{E}}_{x \leftarrow \mathsf{SampNo}(1^n)}[\Delta(x)] \leq \frac{1}{n} + \Pr_{x \leftarrow \mathsf{SampNo}(1^n)}\left[\Delta(x) \geq \frac{1}{n}\right] \tag{54}$$

$$\leq \frac{1}{n} + \mathrm{negl}(n). \tag{55}$$

Hence,

$$\Pr\left[\mathsf{WE.Sec}_{\Pi,\overline{P}}(1^n) = 1\right] \geq 1 - \frac{1}{n} - \mathrm{negl}(n) \geq \frac{3}{4}$$

for all sufficiently large $n$. Thus, $\overline{P}$ is a $\Pi$-adversary with advantage $1/4$ in the sense of Definition 18.

**Oracle access is simulatable.** Fix a $\mathsf{QPT}$ challenger $C$ and reduction $\Sigma$, and let $\Gamma$ be the composition.

By Lemma 24, the output of the phased execution of Definition 22 with the adversary $\overline{P}$ is distributed identically to that of $\Gamma^{\overline{P}}(1^n)$. This phased execution is precisely $\mathsf{DecNo}^{\mathsf{SampNo}}_{\Gamma,S^*}(1^n)$. Indeed, the first stage of $\overline{P}$ samples its instances using $\mathsf{SampNo}$, while its second stage answers every ciphertext using $S^*$.

Similarly, $\Gamma^{P}(1^n)$ is distributed identically to the yes-based experiment $\mathsf{DecYes}^{\mathsf{SampYes}}_{\Gamma,D}(1^n)$. The first stage of $P$ samples each instance together with sufficiently many witness copies using $\mathsf{SampYes}$, and its second stage answers each ciphertext using $\mathsf{WE.Dec}$ with fresh witness copies.

Therefore, (52) is exactly the conclusion of Theorem 26, and the claim follows. □

## 6 Impossibility Result

We are now ready to prove Theorem 2, showing the impossibility of restricted quantum black-box reductions from $\mathsf{WE}$ to falsifiable cryptographic assumptions.

**Theorem 35.** *Let $L$ be a language in $\mathsf{QMA}$ with verification algorithm $Q$ and relation $R_{L,Q}$. Assume that $L$ has a $\mathsf{QMA}$–$\mathsf{QCIP}[2]$ gap problem. Let $\Pi = (\mathsf{WE.Enc}, \mathsf{WE.Dec})$ be a $\mathsf{WE}$ scheme for $R_{L,Q}$. Then, for any falsifiable cryptographic assumption $(C, c)$, one of the following statements holds:*

1. *The assumption $(C, c)$ is false.*
2. *There is no restricted $\mathsf{WE}$–reduction showing the security of $\Pi$ from $(C, c)$.*

*Proof.* Suppose that statement 2 is false. Then there exists a restricted $\mathsf{WE}$–reduction $\Sigma$ establishing the security of $\Pi$ from $(C, c)$.

Apply Definition 19 for black-box reduction with the constant polynomial $p \equiv 4$. There exists a polynomial $p'$ such that, for every $\Pi$-adversary $\overline{P}$ with advantage $1/4$,

$$\Pr\left[1 \leftarrow \langle C(1^n), \Sigma^{\overline{P}}(1^n)\rangle\right] \geq c + \frac{1}{p'(n)} \tag{56}$$

for infinitely many $n \in \mathbb{N}$. Note that $p'$ is fixed independently of the choice of $\overline{P}$.

Set

$$\epsilon(n) := \frac{1}{2p'(n)}.$$

By Theorem 34, there exists an inefficient $\Pi$-adversary $\overline{P}$ with advantage $1/4$ such that oracle access to $\overline{P}$ can be simulated efficiently. In particular, for the restricted reduction $\Sigma$ and challenger $C$, there exists a $\mathsf{QPT}$ algorithm $P$ such that

$$\left|\Pr[\Gamma^{\overline{P}}(1^n) = 1] - \Pr[\Gamma^{P}(1^n) = 1]\right| \leq \epsilon(n) \tag{57}$$

for all sufficiently large $n$, where $\Gamma$ denotes the composed interaction between $C$ and $\Sigma$.

Hence, combining (56) and (57), we obtain, for infinitely many $n$,

$$\Pr\left[1 \leftarrow \langle C(1^n), \Sigma^{P}(1^n)\rangle\right] \geq c + \frac{1}{p'(n)} - \epsilon(n) \tag{58}$$

$$= c + \frac{1}{2p'(n)}. \tag{59}$$

Since both $P$ and $\Sigma$ are $\mathsf{QPT}$, the composed algorithm $\Sigma^P$ is also $\mathsf{QPT}$. Thus, $\Sigma^P$ breaks the falsifiable assumption $(C, c)$ with non-negligible advantage. Therefore the assumption is false, proving statement 1. □

**Acknowledgements.** TM is supported by JST CREST JPMJCR23I3, JST Moonshot R&D JPMJMS2061-5-1-1, JST FOREST, MEXT QLEAP, the Grant-in Aid for Transformative Research Areas (A) 21H05183, and the Grant-in-Aid for Scientific Research (A) No.22H00522.

**AI Disclosure.** We used Claude Fable 5 and ChatGPT 5.6 to assist with writing proofs and texts. The authors verified the correctness and originality of all content including references.

## References


[1] Gorjan Alagic, Zvika Brakerski, Yfke Dulek, and Christian Schaffner. "Impossibility of quantum virtual black-box obfuscation of classical circuits". In: *Annual International Cryptology Conference*. Springer. 2021, pp. 497–525 (cit. on pp. 3, 6).

[2] Gorjan Alagic, Andrew M Childs, Alex B Grilo, and Shih-Han Hung. "Noninteractive classical verification of quantum computation". In: *Theory of cryptography conference*. Springer. 2020, pp. 153–180 (cit. on p. 4).

[3] Gorjan Alagic, Dana Dachman-Soled, Manasi Shingane, and Patrick Struck. "Quantum Black-Box Separations: Succinct Non-Interactive Arguments from Falsifiable Assumptions". In: *IACR Communications in Cryptology* 2.4 (Jan. 8, 2026). ISSN: 3006-5496. DOI: 10.62056/ayfhp2w9p (cit. on pp. 3, 5, 7, 9–11, 18).

[4] Gorjan Alagic and Bill Fefferman. "On quantum obfuscation". In: *arXiv preprint arXiv:1602.01771* (2016) (cit. on pp. 3, 6).

[5] Per Austrin, Hao Chung, Kai-Min Chung, Shiuan Fu, Yao-Ting Lin, and Mohammad Mahmoody. "On the impossibility of key agreements from quantum random oracles". In: *Annual International Cryptology Conference*. Springer. 2022, pp. 165–194 (cit. on p. 6).

[6] Kaniuar Bacho, Alexander Kulpe, Giulio Malavolta, Simon Schmidt, and Michael Walter. "Compiled nonlocal games from any trapdoor claw-free function". In: *Annual International Cryptology Conference*. Springer. 2025, pp. 642–673 (cit. on p. 4).

[7] Boaz Barak, Oded Goldreich, Rusell Impagliazzo, Steven Rudich, Amit Sahai, Salil Vadhan, and Ke Yang. "On the (im) possibility of obfuscating programs". In: *Annual international cryptology conference*. Springer. 2001, pp. 1–18 (cit. on pp. 2, 6).

[8] Mohammed Barhoush, Tomoyuki Morimae, Ryo Nishimaki, and Takashi Yamakawa. "Separating Non-interactive Classical Verification of Quantum Computation from Falsifiable Assumptions". In: *Advances in Cryptology – CRYPTO 2026*. Ed. by Nadia Heninger and Mike Rosulek. Cham: Springer Nature Switzerland, 2026, pp. 279–310 (cit. on pp. 3–5, 7, 9–11, 18).

[9] James Bartusek. "Secure quantum computation with classical communication". In: *Theory of Cryptography Conference*. Springer. 2021, pp. 1–30 (cit. on p. 4).

[10] James Bartusek, Zvika Brakerski, and Vinod Vaikuntanathan. "Quantum state obfuscation from classical oracles". In: *Proceedings of the 56th Annual ACM Symposium on Theory of Computing*. 2024, pp. 1009–1017 (cit. on pp. 3, 4, 8, 15).

[11] James Bartusek, Aparna Gupte, Saachi Mutreja, and Omri Shmueli. *Classical Obfuscation of Quantum Circuits via Publicly-Verifiable QFHE*. Cryptology ePrint Archive, Paper 2026/043. https://eprint.iacr.org/2026/043. 2026 (cit. on p. 3).

[12] James Bartusek, Fuyuki Kitagawa, Ryo Nishimaki, and Takashi Yamakawa. "Obfuscation of pseudo-deterministic quantum circuits". In: *Proceedings of*

*the 55th Annual ACM Symposium on Theory of Computing*. 2023, pp. 1567–1578 (cit. on pp. 3, 4, 8, 15).

[13] James Bartusek and Giulio Malavolta. "Indistinguishability Obfuscation of Null Quantum Circuits and Applications". In: *13th Innovations in Theoretical Computer Science Conference (ITCS 2022)*. Schloss Dagstuhl–Leibniz-Zentrum für Informatik. 2022, pp. 15–1 (cit. on pp. 2, 3, 15, 18, 19, 42).

[14] John Bostanci, Jonas Haferkamp, Chinmay Nirkhe, and Mark Zhandry. "Separating QMA from QCMA with a Classical Oracle". In: *Proceedings of the 58th Annual ACM Symposium on Theory of Computing (STOC)*. arXiv:2511.09551. ACM, 2026 (cit. on p. 8).

[15] Zvika Brakerski, Nico Döttling, Sanjam Garg, and Giulio Malavolta. "Candidate iO from homomorphic encryption schemes". In: *Journal of Cryptology* 36.3 (2023), p. 27 (cit. on p. 7).

[16] Harry Buhrman, François Le Gall, and Jordi Weggemans. *Classical versus quantum queries in quantum PCPs with classical proofs*. 2024. arXiv: 2411.00946 [quant-ph]. URL: https://arxiv.org/abs/2411.00946 (cit. on pp. 4, 16).

[17] Boyang Chen, Andrea Coladangelo, and Or Sattath. "The power of a single Haar random state: constructing and separating quantum pseudorandomness". In: *arXiv preprint arXiv:2404.03295* (2024) (cit. on pp. 6, 19).

[18] Nai-Hui Chia, Kai-Min Chung, and Takashi Yamakawa. "Classical verification of quantum computations with efficient verifier". In: *Theory of Cryptography Conference*. Springer. 2020, pp. 181–206 (cit. on p. 4).

[19] Kai-Min Chung, Yao-Ting Lin, and Mohammad Mahmoody. "Black-box separations for non-interactive classical commitments in a quantum world". In: *Annual International Conference on the Theory and Applications of Cryptographic Techniques*. Springer. 2023, pp. 144–172 (cit. on p. 6).

[20] Andrea Coladangelo and Saachi Mutreja. "On black-box separations of quantum digital signatures from pseudorandom states". In: *Theory of Cryptography Conference*. Springer. 2024, pp. 289–317 (cit. on pp. 6, 19).

[21] Marc Fischlin and Dominique Schröder. "On the Impossibility of Three-Move Blind Signature Schemes". In: *Advances in Cryptology – EUROCRYPT 2010*. Vol. 6110. Lecture Notes in Computer Science. Springer, 2010, pp. 197–215 (cit. on p. 5).

[22] Sanjam Garg, Craig Gentry, Shai Halevi, Mariana Raykova, Amit Sahai, and Brent Waters. "Candidate indistinguishability obfuscation and functional encryption for all circuits". In: *SIAM Journal on Computing* 45.3 (2016), pp. 882–929 (cit. on p. 2).

[23] Sanjam Garg, Craig Gentry, Shai Halevi, and Daniel Wichs. "On the Implausibility of Differing-Inputs Obfuscation and Extractable Witness Encryption with Auxiliary Input". In: *Advances in Cryptology – CRYPTO 2014*. Vol. 8616. Lecture Notes in Computer Science. Springer, 2014, pp. 518–535 (cit. on p. 6).

[24] Sanjam Garg, Craig Gentry, Amit Sahai, and Brent Waters. "Witness encryption and its applications". In: *Proceedings of the forty-fifth annual*

*ACM symposium on Theory of computing*. 2013, pp. 467–476 (cit. on pp. 2, 18).

[25] Romain Gay and Rafael Pass. "Indistinguishability obfuscation from circular security". In: *Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing*. 2021, pp. 736–749 (cit. on p. 7).

[26] Craig Gentry and Daniel Wichs. "Separating succinct non-interactive arguments from all falsifiable assumptions". In: *Proceedings of the forty-third annual ACM symposium on Theory of computing*. 2011, pp. 99–108 (cit. on pp. 2, 3, 5, 7–11, 18).

[27] Alexandru Gheorghiu and Thomas Vidick. "Computationally-secure and composable remote state preparation". In: *2019 IEEE 60th Annual Symposium on Foundations of Computer Science (FOCS)*. IEEE. 2019, pp. 1024–1033 (cit. on p. 4).

[28] Shafi Goldwasser and Guy N. Rothblum. "On Best-Possible Obfuscation". In: *Theory of Cryptography – TCC 2007*. Vol. 4392. Lecture Notes in Computer Science. Springer, 2007, pp. 194–213 (cit. on p. 6).

[29] Sabee Grewal and William Kretschmer. "Unentanglement and Post-Measurement Branching in Quantum Interactive Proofs". In: *arXiv preprint arXiv:2509.15319* (2025) (cit. on pp. 4, 16).

[30] Sam Hopkins, Aayush Jain, and Huijia Lin. "Counterexamples to new circular security assumptions underlying iO". In: *Advances in Cryptology–CRYPTO 2021: 41st Annual International Cryptology Conference, CRYPTO 2021, Virtual Event, August 16–20, 2021, Proceedings, Part II 41*. Springer. 2021, pp. 673–700 (cit. on p. 7).

[31] Akinori Hosoyamada and Takashi Yamakawa. "Finding collisions in a quantum world: quantum black-box separation of collision-resistance and one-wayness". In: *Journal of Cryptology* 37.4 (2024), p. 35 (cit. on p. 6).

[32] Mi-Ying Huang and Er-Cheng Tang. "Obfuscation of Unitary Quantum Programs". In: *2025 IEEE 66th Annual Symposium on Foundations of Computer Science (FOCS)*. arXiv:2507.11970. 2025 (cit. on pp. 3, 4, 8, 15).

[33] Aayush Jain, Huijia Lin, and Amit Sahai. "Indistinguishability obfuscation from well-founded assumptions". In: *Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing*. 2021, pp. 60–73 (cit. on pp. 2, 7).

[34] William Kretschmer. "Quantum Pseudorandomness and Classical Complexity". In: *16th Conference on the Theory of Quantum Computation, Communication and Cryptography*. 2021 (cit. on p. 6).

[35] Chuhan Lu and Nikhil Pappu. *Notions of Quantum Reductions and Impossibility of Statistical NIZK*. Cryptology ePrint Archive, Paper 2024/1847. 2024. URL: https://eprint.iacr.org/2024/1847 (cit. on p. 19).

[36] Urmila Mahadev. "Classical verification of quantum computations". In: *2018 IEEE 59th Annual Symposium on Foundations of Computer Science (FOCS)*. IEEE. 2018, pp. 259–267 (cit. on p. 4).

[37] Mohammad Mahmoody, Ameer Mohammed, Soheil Nematihaji, Rafael Pass, et al. "A note on black-box separations for indistinguishability obfuscation". In: *Cryptology ePrint Archive* (2016) (cit. on p. 6).

[38] Moni Naor. "On cryptographic assumptions and challenges". In: *Annual International Cryptology Conference*. Springer. 2003, pp. 96–109 (cit. on pp. 2, 3, 18).

[39] Anand Natarajan and Tina Zhang. "Bounding the quantum value of compiled nonlocal games: from CHSH to BQP verification". In: *2023 IEEE 64th Annual Symposium on Foundations of Computer Science (FOCS)*. IEEE. 2023, pp. 1342–1348 (cit. on p. 4).

[40] Rafael Pass. "Limits of Provable Security from Standard Assumptions". In: *Proceedings of the 43rd Annual ACM Symposium on Theory of Computing (STOC)*. ACM, 2011, pp. 109–118 (cit. on p. 5).

[41] Amit Sahai and Brent Waters. "How to use indistinguishability obfuscation: deniable encryption, and more". In: *Proceedings of the forty-sixth annual ACM symposium on Theory of computing*. 2014, pp. 475–484 (cit. on p. 2).

[42] Kabir Tomer and Mark Zhandry. "On the Cryptographic Foundations of Interactive Quantum Advantage". In: *arXiv preprint arXiv:2510.05082* (2025) (cit. on p. 19).

[43] Rotem Tsabary. "Candidate witness encryption from lattice techniques". In: *Annual International Cryptology Conference*. Springer. 2022, pp. 535–559 (cit. on pp. 6, 7).

[44] Vinod Vaikuntanathan, Hoeteck Wee, and Daniel Wichs. "Witness Encryption and Null-IO from Evasive LWE". In: *Advances in Cryptology – ASIACRYPT 2022, Part I*. Vol. 13791. Lecture Notes in Computer Science. Springer, 2022, pp. 195–221 (cit. on pp. 6, 7).

[45] Hoeteck Wee and Daniel Wichs. "Candidate obfuscation via oblivious LWE sampling". In: *Advances in Cryptology–EUROCRYPT 2021: 40th Annual International Conference on the Theory and Applications of Cryptographic Techniques, Zagreb, Croatia, October 17–21, 2021, Proceedings, Part III*. Springer. 2021, pp. 127–156 (cit. on p. 7).

[46] Jiayu Zhang. "Classical verification of quantum computations in linear time". In: *2022 IEEE 63rd Annual Symposium on Foundations of Computer Science (FOCS)*. IEEE. 2022, pp. 46–57 (cit. on p. 4).

# A Separating Indistinguishability Obfuscation

We formalize Corollary 3, ruling out black-box reductions for indistinguishability obfuscation of null quantum circuits to falsifiable assumptions. This is a direct consequence of Theorem 35 and Theorem 17.

We first record the structural form of the construction of [13], which builds WE for QMA from null–qIO.

**Proposition 36.** *Let $\mathcal{O} = (\mathsf{Obf}, \mathsf{Eval})$ be a null–qIO scheme (Definition 5) and let $L \in \mathsf{QMA}$ with verification algorithm $Q$ and relation $R_{L,Q}$. The WE scheme for $L$ $\Pi[\mathcal{O}] = (\mathsf{WE.Enc}, \mathsf{WE.Dec})$ of Theorem 17 has the following form: there is a polynomial $t = t(n)$ and a deterministic polynomial-time map $x \mapsto (Q_0^x, Q_1^x)$ sending an instance to descriptions of two quantum circuits, such that*

1. $\mathsf{WE.Enc}(1^n, x, b) = \mathsf{Obf}(1^n, Q_b^x)$ *for* $b \in \{0,1\}$*;*
2. $\mathsf{WE.Dec}(x, \mathsf{ct}, \rho) = \mathsf{Eval}(\mathsf{ct}, \rho)$*;*
3. $Q_0^x$ *is the null circuit for every* $x$*, and* $Q_1^x$ *is a null circuit whenever* $x \notin L$*.*
4. *for every* $x \in L$ *with witness* $|w_x\rangle$*,* $\Pr[Q_1^x(|w_x\rangle^{\otimes t}) = 1] \geq 1 - \mathsf{negl}(n)$*.*

Here $Q_1^x$ is the soundness-amplified verifier $Q(x, \cdot)$: running $Q$ in parallel on independent witness copies and thresholding drives the QMA soundness error below any negligible function, so that $Q_1^x$ rejects all inputs except with negligible probability when $x \notin L$ (item 3), while still accepting the multi-copy witness when $x \in L$ (item 4). The circuits are padded to have the same size and input/output lengths. Items 1–4 recover the correctness (Definition 15) and security (Definition 16) of $\Pi[\mathcal{O}]$: decryption evaluates the obfuscation on the witness, and for $x \notin L$ both $Q_0^x, Q_1^x$ are null, so the two obfuscations are indistinguishable by null–qIO security.

The security notion for null–qIO in [13] (Definition 8) is defined relative to a non-interactive non-uniform adversary. We present a weaker definition here mirroring Definition 16. As in the WE game, the adversary is permitted to choose the challenge itself (here, the pair of null circuits) and we only consider uniform adversaries. Both changes strengthen the resulting separation. Note that it is straightforward to show that the WE construction from null–qIO holds relative to this weakened security notion.

**Definition 37 (Null–qIO security game).** *Let $\mathcal{O} = (\mathsf{Obf}, \mathsf{Eval})$ be a null–qIO scheme and $A = (A_1, A_2)$ be a pair of quantum algorithms.*

$\mathsf{qIO.Sec}_{\mathcal{O},A}(1^n)$*:*

1. $b \leftarrow \{0,1\}$.
2. $(Q_0, Q_1, z) \leftarrow A_1(1^n)$*, where* $z$ *is some classical string.*
3. *If the circuits* $(Q_0, Q_1)$ *are invalid (do not have the same size and the same input and output spaces) output* $0$*. Otherwise, compute* $\widetilde{Q} \leftarrow \mathsf{Obf}(1^n, Q_b)$*.*
4. $b' \leftarrow A_2(Q_0, Q_1, z, \widetilde{Q})$.
5. *Output* $1$ *if* $Q_0$ *and* $Q_1$ *are both null circuits and* $b' = b$*; otherwise output* $0$*.*

*Remark 38.* Note that in Definition 37 the first phase of the adversary $A_1$ can pass a classical string $z$ to the second phase $A_2$. This is necessary to mirror the $\mathsf{WE}$ game where the instance $x$ is sent from the first phase to the second phase. Note that Definition 37 is still weaker than the standard security notion (Definition 8), which allows for quantum advice.

**Definition 39 (qIO–adversary).** *A quantum algorithm $A = (A_1, A_2)$ is a* qIO–adversary *for $\mathcal{O}$ if there is a polynomial $p$ such that $\Pr[\mathsf{qIO.Sec}_{\mathcal{O},A}(1^n) = 1] \geq \frac{1}{2} + \frac{1}{p(n)}$ for infinitely many $n$.*

**Definition 40 (Null–qIO–reduction).** *Let $(C, c)$ be a falsifiable assumption. A quantum black-box reduction showing the security of $\mathcal{O}$ from $(C, c)$ is a $\mathsf{QPT}$ algorithm $\Sigma^{(\cdot)}$ such that for every polynomial $p$ there is a polynomial $p'$ with the following property: for every qIO–adversary $A$ of advantage $1/p$, the algorithm $\Sigma^A$ breaks $(C, c)$ with advantage $1/p'$ i.e.*

$$\Pr\left[1 \leftarrow \langle C(1^n), \Sigma^A(1^n)\rangle\right] \;\geq\; c + \frac{1}{p'(n)}, \tag{60}$$

*for infinitely many $n \in \mathbb{N}$. For short, we call such a reduction a* null–qIO–reduction.

Our separation applies to reductions with the following structure.

**Definition 41 (Restricted null–qIO reduction).** *Let $\mathcal{O} = (\mathsf{Obf}, \mathsf{Eval})$ be a null–qIO scheme, let $(C, c)$ be a falsifiable cryptographic assumption, and let $\Sigma^{(\cdot)}$ be a null–qIO reduction showing the security of $\mathcal{O}$ from $(C, c)$.*

*We say that $\Sigma^{(\cdot)}$ is a* restricted null–qIO reduction *if there exist polynomials $s = s(n)$ and $\ell = \ell(n)$ such that, in the interaction*

$$\left\langle C(1^n), \Sigma^A(1^n)\right\rangle,$$

*the oracle access of $\Sigma$ to a possibly inefficient null–qIO adversary $A = (A_1, A_2)$ has the following form.*

1. *The challenger $C$ and the reduction $\Sigma$ may interact arbitrarily throughout the execution. Only $\Sigma$ has oracle access to $A$.*
2. *For every $i \in [s]$, the reduction submits $1^n$ to $A_1$ and receives a syntactically valid pair of quantum-circuit descriptions*

   $$(Q_i^0, Q_i^1, z_i) \leftarrow A_1(1^n).$$

   *where $z_i$ is a polynomial-length classical string and $Q_i^0, Q_i^1$ are syntactically valid circuit descriptions having the same size and the same input and output spaces.*
3. *For every $i \in [s]$ and $j \in [\ell]$, a bit $b_{i,j} \leftarrow \{0,1\}$ is sampled, and*

   $$\widetilde{Q}_{i,j} \leftarrow \mathsf{Obf}\left(1^n, Q_i^{b_{i,j}}\right)$$

   *is generated honestly. This computation may be performed within $\Sigma$, within $C$, or jointly through their interaction.*

4. *For every $i \in [s]$ and $j \in [\ell]$, the reduction submits $(Q_i^0, Q_i^1, z_i, \widetilde{Q}_{i,j})$, and receives*
$$\widetilde{b}_{i,j} \leftarrow A_2\Big(Q_i^0, Q_i^1, z_i, \widetilde{Q}_{i,j}\Big).$$
5. *The execution ends when $C$ outputs its decision bit.*

*Thus, the oracle queries are parameter-preserving, classical, non-adaptive, and made only on honestly generated obfuscations.*

The following directly follows from Proposition 36 and Theorem 35.

**Theorem 42.** *Let $L$ be a language in* $\mathsf{QMA}$ *admitting a* $\mathsf{QMA}$–$\mathsf{QCIP}[2]$ *gap problem, and let $\mathcal{O} = (\mathsf{Obf}, \mathsf{Eval})$ be a null–*$\mathsf{qIO}$ *scheme. Then for every falsifiable assumption $(C, c)$, one of the following holds:*

1. *the assumption $(C, c)$ is false; or*
2. *there is no restricted null–*$\mathsf{qIO}$*–reduction showing the security of $\mathcal{O}$ from $(C, c)$.*