PhantomCrypt: Composing Existence and Content Deniability under a Post-Quantum Envelope

Shahzad Ahmad[1][0000-0002-9654-869X], Stefan Rass[1][0000-0003-2821-2489], and Zahra Seyedi[2][0009-0002-8492-4640]

[1] LIT Secure and Correct Systems Lab, Johannes Kepler University, Linz, Austria
[2] Computer Science and Engineering Department, Koç University, Istanbul, Türkiye,
shahzad.ahmad@jku.at, stefan.rass@jku.at, zseyedi@ku.edu.tr

Note. This is the full version. A condensed version of this work, containing the definitions, the construction, the ED proof, the CD lemmas in sketch, and the feasibility threshold, is under submission to a peer-reviewed venue. This version additionally contains the complete CD proofs, the open-world (t, q) framework, multi-message security, the seed-management analysis, and the extended related-work treatment.

Abstract. Traditional deniable encryption denies the *content* of secret communications by allowing plausible alternative plaintexts under coercion. But the recognizable use of a deniable-encryption tool can itself defeat the purpose: a revealed plaintext becomes suspicious once a coercer detects that a non-standard tool was used, and fully encrypted, format-anomalous traffic is already detected and acted upon by deployed censorship systems. This motivates a second axis of deniability, the deniability of the mechanism's *use*, which we call *second-order deniability* (2OD) and treat as a design goal rather than a single achievable primitive. We decompose 2OD into two formally separate properties: *content deniability* (CD), that a coercer holding all disclosed keys cannot identify the true message among decoys, and *existence deniability* (ED), that the transmitted ciphertext is indistinguishable from the output of a fixed, standard reference protocol. We give game-based definitions of CD and ED, prove that the two are logically independent, and present PhantomCrypt, a construction that composes False-Bottom Encryption (CD) with Invisible Encryption (ED) under a post-quantum hybrid KEM/AEAD envelope. We are deliberate about what each guarantee buys. ED, as we define and prove it, is the indistinguishability of the transmitted ciphertext *object* from a single, explicitly specified reference distribution (a standard hybrid KEM/AEAD wrapper over an unmodified cover string); it reduces to KEM and AEAD security. It is *not* a claim of traffic-analysis resistance, and it is conditional on an assumption we make explicit rather than prove: that the chosen cover string is itself unremarkable to the observer. CD we prove negligible in the random-oracle model, conditioned on the secrecy of a pre-shared seed, with the residual advantage controlled by two explicit quantities: the seed's entropy relative to the field size, and a cover-concentration term that we introduce here and

bound by a min-entropy admissibility condition validated on real text. Post-quantum security enters through the envelope only: ED reduces to ML-KEM and AEAD security against a quantum distinguisher, whereas the CD argument is statistical and stated in the classical random-oracle model, so we claim a post-quantum *envelope* rather than post-quantum deniability wholesale and leave a QROM treatment of composed CD open. We further show that this coupling is sharp: a feasibility threshold separates a regime where strong deniability holds from one where it is information-theoretically impossible, governed by the seed-entropy margin relative to the plausible-message entropy. We are explicit about the boundary throughout: ED is object-level indistinguishability, not a defense against traffic analysis or device-side software forensics, and we state where full 2OD does and does not hold. A proof-of-concept implementation encrypts a 32-byte message with three decoys in under 10 ms on commodity hardware, with scaling measured across cover-text length and decoy count.

Keywords: Deniable encryption · Plausible deniability · Post-quantum cryptography · Secret sharing · Steganography

1 Introduction

The growing sophistication of state-sponsored surveillance and coercive interrogation poses significant challenges to privacy. Traditional cryptographic solutions provide confidentiality, but they fail under coercion: if an adversary forces disclosure of decryption keys, the protected content is exposed. Classical deniable encryption, introduced by Canetti et al. [12], addresses this by enabling parties under duress to present a plausible but false message. We call this property *content deniability* (CD).

However, a critical limitation remains. Specialized deniable encryption schemes produce ciphertexts that do not conform to any standard protocol format. An observer who detects data in an unrecognized format can conclude that non-standard encryption is being used. This is not hypothetical: Wu et al. [42] document that the Great Firewall of China actively detects and blocks fully encrypted traffic precisely because its byte distribution fails to match any whitelisted protocol, and the Fully Encrypted Protocol line of work [24, 25] formalizes exactly the property that censors exploit. In such settings, the detection of format-anomalous or unexplained encryption is itself treated as grounds for blocking, escalation, or suspicion, independent of whether the content is ever recovered.

This gap motivates *existence deniability* (ED): the property that covert communication is indistinguishable from innocuous data, including legitimate encrypted traffic. Where CD asks "which message was encrypted?", ED asks "was anything encrypted at all?"

We state at the outset what our ED guarantee does and does not deliver, because the motivation above concerns live channels, whereas our provable guarantee concerns the ciphertext *object*. We prove that a PhantomCrypt ciphertext

is indistinguishable from a single, explicitly fixed reference distribution: a standard hybrid KEM/AEAD wrapper carrying payloads of the matching length, with the (unmodified) cover string alongside. This reduces to KEM and AEAD security. It does *not* by itself defeat a censor that keys on traffic shape, timing, volume, or the plausibility of the cover string, and it is conditional on an assumption we make explicit rather than discharge: that the chosen cover string is unremarkable to the observer. The detection results of Wu et al. [42] target byte distributions of the transmitted object, which is precisely what our ED game models; the residual traffic-analysis surface is real, and we treat it as out of scope (Remark 2) and as an open problem (Section 9), not as something the present proofs close.

We use the term *second-order deniability* (2OD) for the design goal that the use of a deniability mechanism is itself deniable. First-order deniability protects the *content* of a message (CD); second-order deniability additionally targets the *existence* of the deniable communication (ED). We treat 2OD as a goal decomposed into two formally separate, provable properties rather than a single primitive with a single security game, and we are precise throughout about the regime in which it holds. In particular, 2OD in the strong sense "even an adversary inspecting the device cannot tell the mechanism was used" is *not* what we prove; we prove network-level ED and coercion-time CD, and we delimit the gap explicitly in Section 1.1.

Contributions. We formalize CD and ED via game-based definitions that are agnostic to any specific construction, and prove that they are logically independent. We characterize 2OD as the conjunction of the two, stating this characterisation at the level it supports rather than as a security game with a reduction (Section 3.4). We present PhantomCrypt, composing FBE (for CD) with IE (for ED) within a hybrid PQC-KEM envelope, prove ED by reduction to KEM and AEAD security against a concrete reference distribution, and prove CD negligible in the random oracle model, conditioned on the secrecy of a pre-shared seed and with the residual advantage controlled by the seed entropy relative to the field size (Theorem 4). We give security as an explicit function of the secret-sharing threshold k, identify the regime in which the open-world (t, q) guarantee holds, and benchmark at the k that the security analysis actually requires rather than at a smaller k. Our main analytic contribution is a *feasibility threshold* (Section 7): undetectability and strong identical-distribution coercion deniability are not independently dialable but are coupled through a single seed-entropy margin $s = \lambda - H(\mathcal{P})$ relative to the entropy of the plausible-message set. We prove a coverage law $1 - \exp(-2^s)$, a feasibility direction above the threshold, and a matching *information-theoretic impossibility* below it: when the margin is negative, a $1 - o(1)$ fraction of freely chosen fakes cannot be opened by any seed, so strong deniability fails against even an unbounded adversary. Detection-only analyses of the underlying scheme never encounter this line, since they do not require fakes to be openable; it is the precise price of adding the second deniability axis. We identify key-reuse vulnerabilities with mitigations, demonstrate practical feasibility through measured scaling, and are explicit about the threat

model boundary: ED is indistinguishability of the ciphertext object from a fixed reference protocol, not resistance to traffic analysis or device-side forensics.

1.1 Threat Model: A Two-Stage Adversary

A natural objection is that CD and ED are redundant: if ED succeeds, no coercer suspects anything, making CD unnecessary; if ED fails, CD is the only defense. We address this with a two-stage threat model. In Stage 1 (network observation), the adversary passively observes traffic; ED protects here. If ED holds, the adversary has no cryptographic reason to escalate. In Stage 2 (device seizure or coercion), the adversary escalates for reasons *outside* our model (an informant, an unrelated investigation), seizes the device, and coerces disclosure of keys. The two properties play distinct roles, and we are deliberate about which survives Stage 2.

What the coerced party actually does. A recurring concern with deniable encryption is that *refusing* to surrender a key is itself incriminating. PhantomCrypt is designed so that such a refusal cannot occur. Under coercion, the party surrenders the device and all $\ell+1$ decoy keys, each of which carries the on-device KEM key inside it, and asserts that the transcript is an ordinary hybrid-encrypted message. This assertion is consistent with everything the adversary can check: the disclosed keys open the ciphertext to $\ell+1$ plausible messages, none distinguished as true (CD), and the ciphertext object matches the reference protocol (ED). The deniable secret sk_{den} is a pre-shared seed that is never stored on the device and whose existence is not evidenced by either the ciphertext or the disclosed keys. There is therefore no observable "withheld key": the adversary is given a complete, self-consistent account, and the residual secret leaves no gap in that account that the adversary can point to. This is what distinguishes the design from naively pre-sharing one real and one fake credential, where the adversary knows exactly two openings exist and that one is being suppressed.

Two objections to this picture deserve a direct answer, since either, if unanswered, would make CD pointless. The first is that surrendering $\ell+1$ keys is itself evidence that the user is "playing a game", which may suffice for punitive action regardless of content. Three things bear on this. Structurally, the disclosed material is one container of field elements together with $\ell+1$ keys of identical format, which is the shape of an ordinary multi-record encrypted archive with per-record keys; nothing in the tuple is specific to deniable encryption. Procedurally, every opening is a call to the same Dec with the same arity (Definition 1), so the coercer has no procedure left to invoke and no argument pattern to ask about, which is precisely why the syntax exposes one decryption algorithm rather than two. Making the structural claim formal against a deployed container format is the *conformance* requirement we identify but do not discharge (Section 9). Beyond all three, cryptography cannot force a coercer to behave rationally: an adversary willing to punish mere possession of encrypted data defeats every encryption scheme and lies outside any cryptographic model. What CD guarantees is the weaker but well-defined property that the coercer gains no *evidence*

about which opening is true, so punishment must be indiscriminate rather than targeted at m^*.

The second objection is that m^* is itself among the $\ell+1$ recovered openings, so the user can be punished for its content. This is correct, and we state it as a design property rather than hide it: CD conceals the true message's *designation*, not its content (Remark 7). The scheme is meaningful exactly when m^* is a plausible member of the decoy set, for instance when all $\ell + 1$ payloads are high-entropy values (keys, tokens, hashes) of which any could be innocuous. When m^* is self-incriminating on its face, no decoy set repairs that, and the appropriate tool is the anamorphic model discussed in Section 4. Decoy credibility is a deployment responsibility outside the cryptographic model, consistent with empirical findings that the credibility of a denial, rather than its cryptographic soundness, often decides outcomes [32].

Where 2OD holds and where it does not. 2OD in the full sense, that an adversary cannot tell a deniability mechanism was *ever* used, requires both the ciphertext and the *tool that produced it* to be deniable. We achieve the former (ED) and not the latter. If forensic inspection of a seized device reveals PhantomCrypt software, the adversary learns that a deniability mechanism is available to the user, and 2OD in the full sense no longer holds; CD still does, so the true message is not identified. Software-forensic deniability is a deployment problem (e.g., shipping the mechanism inside a tool whose presence is unremarkable, such as a standard messaging or TLS library), distinct from the cryptographic question this paper addresses, and we do not claim to solve it. We flag it as the principal open direction rather than burying it: the conjunction CD $\wedge$ ED provides defense-in-depth at the network and coercion layers, and the device layer is out of scope.

2 Conceptual Overview

PhantomCrypt targets second-order deniability by building on a single cryptographic primitive: threshold secret sharing via Lagrange interpolation over $\mathbb{F}_p$, whose perfect secrecy is information-theoretic. This primitive underpins both Invisible Encryption (IE) for existence deniability and False-Bottom Encryption (FBE) for content deniability. We explain the core idea through a concrete scenario before formalizing it.

The scenario. Sarah wishes to send a secret message m_1 to Robert. She composes a natural-language cover text $T = (w_1, \ldots, w_L)$ for instance, an ordinary email or article. Using a pre-shared seed x_0, Sarah produces a pseudorandom sequence $x_1, \ldots, x_{k-1}$ by cryptographic means, and, with another pseudorandom sequence, selects $k-1$ words $w_{i_1}, \ldots, w_{i_{k-1}}$ from the text. Each selected word defines a share $s_j \leftarrow H(w_{i_j}) \in \mathbb{F}_p$, where H is a hash function modeled as a random oracle. These shares, together with the message m_1 placed at the

secret point $P(0)$, determine a unique polynomial P of degree at most $k-1$ via Lagrange interpolation through the k points,

$$(0,\, m_1),\ (x_1,\, s_1),\ \ldots,\ (x_{k-1},\, s_{k-1}). \tag{1}$$

Sarah evaluates P at a fresh abscissa x_{new} (likewise pseudorandomly derived from x_0) to obtain the transmitted share $s_{\text{new}} = P(x_{\text{new}})$. By the security of Shamir secret sharing, s_{new} on its own reveals nothing about m_1: it is a uniformly random element of $\mathbb{F}_p$. Because it looks random, s_{new} can be carried unsuspiciously as the payload wrapped inside a standard hybrid encryption envelope (a Key Encapsulation Mechanism followed by authenticated symmetric encryption), where it is indistinguishable from the random-looking data such an envelope normally transports. We fix this single carrier throughout: s_{new} is an AEAD plaintext under a KEM-derived session key, never the session key itself. The cover text is transmitted verbatim: no word is modified, and the shares are *derived from* it rather than embedded into it. This is the Invisible Encryption mechanism [3], and it provides existence deniability.

Adding content deniability. Suppose Sarah is under coercion and must explain her transcript. With IE alone, she can claim the transcript is an ordinary hybrid-encrypted message. Without the seed x_0 or knowledge of which words serve as shares, the coercer cannot recover m_1. However, if the coercer becomes suspicious and Sarah's device is compromised, there is only one message to find.

Sarah has already foreseen this possibility. Before coercion, she has already embedded a second message m_2 using a different set of pseudorandomly selected words (shown with gray highlighting in Figure 1), each producing its own share. Figure 1 illustrates a concrete transcript showing both mechanisms: the final share for m_2 is carried as the AEAD-wrapped payload of the hybrid envelope, and the final share for m_1 is carried in a QR code disguised as a login token.

A recent scheme called False-Bottom Encryption (FBE) [1] formalizes this multi-message capability: it embeds an arbitrary number of messages within a single ciphertext via underdetermined linear systems over $\mathbb{F}_p$.

The combination. IE and FBE are combinable because they rest on the same mechanism: secret sharing over $\mathbb{F}_p$. To make IE content-deniable, Sarah embeds m^* and ℓ decoy messages into an FBE container c_{FBE}; the IE mechanism then derives shares from the cover text and computes a final share s_{new} encoding m^*. This share can be delivered in two ways: (a) as the AEAD-wrapped payload of an ordinary hybrid encryption envelope, or more generally as random-looking material in any protocol field that legitimately carries random bytes[3], in which case the cover text remains unmodified; or (b) steganographically embedded in the cover text (QR code or image), in which case the cover grows with each added message as in standalone FBE.

[3] Protocols whose randomness is not made visible to the opposite party (e.g., probabilistic encryption, certain signature schemes) are not suitable carriers for s_{new}.

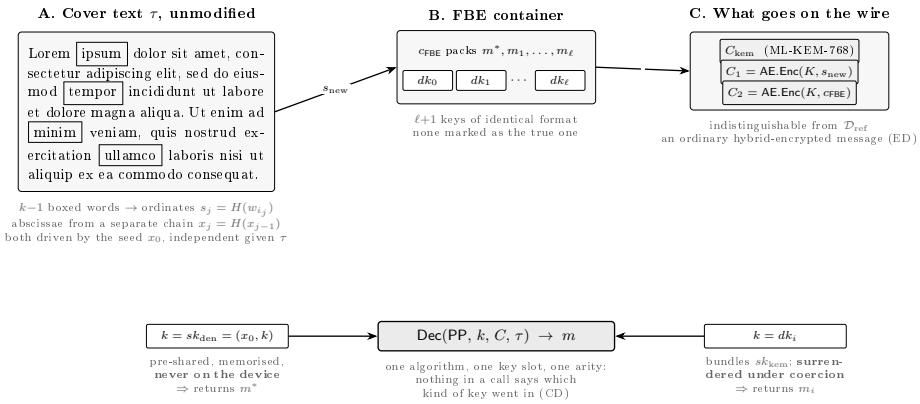

Fig. 1: PhantomCrypt at a glance. The cover text carries shares for the true message m^* only (A); the decoys live in the FBE container, which exposes $\ell + 1$ structurally identical keys (B); the transmitted object is an ordinary hybrid-encrypted message (C). Every opening, truthful or decoy, is the same call to Dec with one key slot, so the interface itself does not distinguish them. Only sk_{den} is withheld under coercion, and it was never an object on the device.

Robert, who shares the seed x_0, reconstructs m^* by re-deriving shares and interpolating. Under coercion the openings come from the FBE container, not from the IE layer: the coerced party surrenders the decoy keys and each one opens the container to a different plausible message. Reinterpolating a different polynomial through the same shares is *not* the mechanism, and would not survive the CD game, because such an object is not a key the scheme ever emitted (Remark 6). The deniability mechanism is itself deniable: an adversary without x_0 cannot tell that s_{new} served as a share, let alone which message it encodes. FBE provides CD, IE provides ED, and because both rest on the same algebraic foundation, the use of either mechanism is deniable.

3 Formal Definitions

We formalize content deniability, existence deniability, and second-order deniability through game-based definitions. All definitions are agnostic to the choice of building blocks; the specific construction PhantomCrypt is introduced in Section 5. Table 1 lists the principal symbols. For a string x, we write $|x|$ for its length, and $\text{poly}(\cdot)$ will denote any arbitrary but fixed univariate polynomial.

3.1 Scheme Syntax

Our scheme is a *symmetric plan-ahead* type of deniable encryption (as opposed to ad hoc or public-key schemes, whose construction we leave for future work). Before the syntax, one word about *devices*, because everything under coercion

Table 1: Notation: symbols and their meanings.

Symbol	Type	Meaning
λ	Integer	Security parameter
p	Prime	Defines $\mathbb{F}_p$
ℓ	Integer	Number of decoy messages
τ / T	String	Cover string (generic) / cover text (PhantomCrypt)
sk_{den}	Key	Deniable secret key (pre-shared, not in disclosed material)
dk_i	Key	i-th decoy key, $dk_i = (sk_{\mathrm{kem}}, \mathsf{SK}_{\mathsf{FBE},i})$
m^*, m_i	$\{0,1\}^{\mathrm{poly}(\lambda)}$	True message; i-th decoy message
k	Integer	Secret-sharing threshold of the IE layer
t, q	Integers	Parameters of the open-world (t, q) notion; t is *not* the sharing threshold
$\mathsf{SK}_{\mathrm{ie}}$	Key tuple	IE secret (x_0, k); instantiates sk_{den}
sk_{kem}	Key	KEM decapsulation key; on-device, carried inside every dk_i
$\mathsf{SK}_{\mathsf{FBE},i}$	Key	i-th FBE key; the second component of dk_i
$f_{\max}(\tau)$	$[0,1]$	Largest fraction of cover positions sharing one share value
$H_\infty(\tau)$	Bits	Share min-entropy $-\log_2 f_{\max}(\tau)$ of one selection
s_j, s_{new}	$\mathbb{F}_p$	j-th share $H(w_{i_j})$; transmitted share $P(x_{\mathrm{new}})$
c_{FBE}	Vector	FBE ciphertext container
C_{kem}, C_1, C_2	Bytes	KEM ciphertext; AEAD-encrypted payloads
H	Hash	$\{0,1\}^* \to \mathbb{F}_p$ (random oracle)

turns on where key material physically sits. We model the coerced party as owning a single device, a phone or a laptop, which holds the files, the software, and the key material needed for ordinary operation. A coercer who reaches the party reaches the device and can compel every secret the device holds. Keys the party cannot produce on demand must therefore live somewhere the device does not reach: a short memorised passphrase or mnemonic, a secret pre-shared out of band and held only on the recipient's side, or a secret derived on the fly from something the party knows. None of these can be an arbitrarily long string, so the design places exactly one small, memorable secret off the device and lets everything bulky sit on it.

Compared to the classical three-algorithm syntax of [12], which has a single key and cannot express the difference between a secret that lives on the device and a secret that was never written there, we separate two key roles: the long-lived *deniable secret* sk_{den}, produced once by setup and never written to the device, and the per-message *decoy keys* $\{dk_i\}$ emitted by encryption, one for each message a coerced party may later choose to open. Both kinds are accepted by the same decryption algorithm.

Why exactly one decryption algorithm. Keeping decryption single is a design constraint rather than a convenience, and it is worth saying why before the definition rather than after it. Suppose the syntax offered two procedures, one for the true message and one for a decoy. The specification would then name a true opening, and a coercer who reads the specification could demand that the party run it. The party would have to refuse, and a refusal is an admission. Worse, the two procedures would take different arguments, so the very shape of a call would reveal which kind of opening was being performed. A scheme whose interface distinguishes truthful decryption from decoy decryption has given the mechanism away in its own documentation. The syntax below therefore exposes exactly one decryption algorithm, with one key slot and one arity. Which message comes out depends only on which key goes in, and no observer of the interface can tell the two cases apart.

Why not simply pre-share a real and a fake key? The setup assumes a pre-shared symmetric secret, so one might ask why the secret-sharing machinery is needed at all rather than pre-distributing a single real credential and a single fake one. Two reasons. First, a real/fake credential pair yields exactly two openings and tells the adversary that exactly two exist, so once both are produced the residual uncertainty is zero; FBE yields $\ell+1$ structurally symmetric openings drawn from a super-polynomial pool of candidate decoy sets, so a polynomial-time coercer cannot even establish that no further message is hidden (the (t, q) notion below). Second, a fake credential is a distinct object that the adversary can demand and whose suppression is observable, whereas the secret-sharing route lets the coerced party disclose a complete, self-consistent key set with no missing element to point at (Example 1). The machinery buys symmetric multiplicity and the unmodified-cover-text format that a credential pair cannot provide.

Definition 1 (Deniable Encryption with Existential Hiding (DEEH)). *A* deniable encryption scheme with existential hiding *over a message space $\mathcal{M}$ is a tuple of PPT algorithms*

$$\Pi = (\mathsf{Setup},\ \mathsf{Enc},\ \mathsf{AddDecoy},\ \mathsf{Dec}).$$

The decoy budget $\ell \in \mathbb{N}$*, fixed at setup, is the maximum number of* AddDecoy *invocations the scheme supports. Together with the true message it determines the $\ell+1$ openings available under coercion. Two conditions tie the budget to the cover string: the budget is polynomially bounded, $\ell \leq poly(\lambda)$, and the cover is long enough to carry it, $|\tau| \in \Omega(\ell)$. A cover string meeting both is called* admissible *for Π, and every guarantee below is claimed only over admissible covers. The second condition needs a word, because its force differs between constructions. A scheme that writes its decoys into the cover itself cannot escape it: asking for a thousand decoys inside a one-bit cover is not a hard problem but an impossible one, and the requirement is then a statement about capacity. PhantomCrypt carries its decoys in a separate algebraic container rather than in the cover, so for that construction the condition is forced by plausibility instead: the ciphertext grows with the decoy budget, and a very short cover carrying a very large encrypted payload is exactly the kind of object a passive observer would find odd.*

- $\mathsf{Setup}(\lambda, \tau, \ell) \to (\mathsf{PP}, sk_{\mathrm{den}})$.
  Outputs public parameters PP *and a* deniable secret sk_{den}*. The deniable secret is pre-shared between the communicating parties and is never stored on the device. It is short enough to be memorised or pre-shared out of band, which is what makes withholding it credible under coercion. Any bulk key material a deployment needs on the device, a decapsulation key for instance, is carried inside the keys the scheme emits rather than named separately here, so that the syntax stays independent of how a construction builds its envelope.*
- $\mathsf{Enc}(\mathsf{PP}, sk_{\mathrm{den}}, m^*, \tau) \to (C, dk_0)$.
  Encrypts the true message $m^* \in \mathcal{M}$ under the cover string τ and outputs the ciphertext C together with a decoy key dk_0 whose decoy opening also yields m^*. The requirement that dk_0 be indistinguishable from keys later produced by AddDecoy *is a security property, made precise in the content-deniability game below, not part of the syntax.*
- $\mathsf{AddDecoy}(\mathsf{PP}, sk_{\mathrm{den}}, C, m_i) \to (C', dk_i)$.
  Embeds a decoy message $m_i \in \mathcal{M}$ into the given ciphertext C, for $1 \leq i \leq \ell$, and outputs the updated ciphertext C' together with the decoy key dk_i that opens it. Adding a decoy requires the deniable secret, because the update must leave the true embedding intact. The requirement has a familiar shape: a user of an encrypted volume with a hidden partition must supply the hidden-volume password when writing to the outer volume, not because the outer write needs it, but because without it the write would overwrite the hidden data.
- $\mathsf{Dec}(\mathsf{PP}, k, C, \tau) \to m$.
  The single decryption algorithm. Its key slot accepts any element of the key space, which contains the deniable secret and every decoy key the scheme has emitted. On $k = sk_{\mathrm{den}}$ it returns the true message, and on $k = dk_i$ it returns

the message that key was issued for. The call is the same in both cases, and nothing in the arguments records which kind of key was supplied.

Example 1 (The three key roles under coercion). Sarah runs PhantomCrypt with $\ell = 4$. Setup gives her $sk_{\mathrm{den}} = (x_0, k)$, a 24-word mnemonic she memorises and never writes down. Her phone, the device a coercer would seize, holds an ML-KEM decapsulation key, which travels inside each decoy key rather than standing on its own. She encrypts the true message and adds four decoys. That leaves five keys, $dk_0, \ldots, dk_4$, one more than the decoy count, because Enc emits dk_0 for the true message alongside the four keys $\mathsf{AddDecoy}$ emits for the decoys. Coerced, she surrenders the phone and all five decoy keys. Feeding each one to Dec yields a different, syntactically valid message, and every call looks identical from the outside. The coercer's inventory is complete: there is no key he can name that she has withheld, because sk_{den} never existed as an object on the device. That is the difference between this syntax and a real-plus-fake credential pair, where the coercer knows exactly two openings exist and that one is being suppressed.

Definition 2 (Correctness). *A DEEH scheme is* correct *if, for all $\lambda \in \mathbb{N}$, all admissible cover strings τ, all decoy budgets ℓ, and all pairwise distinct $m^*, m_1, \ldots, m_\ell$, the following holds with probability 1 over the coins of all algorithms. Let*

$$\begin{aligned}(\mathsf{PP}, sk_{\mathrm{den}}) &\leftarrow \mathsf{Setup}(\lambda, \tau, \ell),\\ (C_0, dk_0) &\leftarrow \mathsf{Enc}(\mathsf{PP}, sk_{\mathrm{den}}, m^*, \tau),\end{aligned}$$

and iteratively $(C_i, dk_i) \leftarrow \mathsf{AddDecoy}(\mathsf{PP}, sk_{\mathrm{den}}, C_{i-1}, m_i)$ for $i = 1, \ldots, \ell$. Then (i) $\mathsf{Dec}(\mathsf{PP}, sk_{\mathrm{den}}, C_\ell, \tau) = m^*$; (ii) $\mathsf{Dec}(\mathsf{PP}, dk_0, C_\ell, \tau) = m^*$; and (iii) $\mathsf{Dec}(\mathsf{PP}, dk_i, C_\ell, \tau) = m_i$ for every $i \in \{1, \ldots, \ell\}$. The true message therefore has two openings, one through the deniable secret and one through an ordinary decoy key, and only the first of them is withheld under coercion.

3.2 Content Deniability (CD)

Content deniability captures the property that an adversary who coerces the honest party into disclosing decryption keys cannot identify which message was originally designated as the true one among the decoys. We formalize this via an indistinguishability game, then introduce a complementary asymptotic notion capturing the adversary's *residual uncertainty* about how many messages are hidden.

Definition 3 (CD Security Game). *For a DEEH scheme Π with decoy budget ℓ and adversary $\mathcal{A}$, the game $\mathsf{Game}^{\mathrm{CD}}_{\Pi,\mathcal{A}}(\lambda)$ proceeds as follows:*

1. *$\mathcal{A}$ submits an admissible cover string τ (Definition 1) and $\ell + 1$ pairwise distinct messages $m_0, m_1, \ldots, m_\ell$.*
2. *The challenger runs $(\mathsf{PP}, sk_{\mathrm{den}}) \leftarrow \mathsf{Setup}(\lambda, \tau, \ell)$.*

3. *The challenger samples* $b^* \xleftarrow{\$} \{0,\dots,\ell\}$*, sets* $m^* := m_{b^*}$*, and runs* $(C, dk_{b^*}) \leftarrow \mathsf{Enc}(\mathsf{PP}, sk_{\mathrm{den}}, m^*, \tau)$*. Then, for every* $i \in \{0,\dots,\ell\} \setminus \{b^*\}$*, the challenger runs* $(C, dk_i) \leftarrow \mathsf{AddDecoy}(\mathsf{PP}, sk_{\mathrm{den}}, C, m_i)$*.*
4. *The challenger hands* $\mathcal{A}$ *the tuple* $(\mathsf{PP}, \tau, C, \{dk_i\}_{i=0}^{\ell})$*, so* $\mathcal{A}$ *may run* Dec *on every disclosed key and read all* $\ell+1$ *messages. Only the deniable secret* sk_{den} *is withheld.*
5. $\mathcal{A}$ *outputs a guess* $b' \in \{0,\dots,\ell\}$*;* $\mathcal{A}$ *wins iff* $b' = b^*$*.*

Define $\mathsf{Adv}^{\mathrm{CD}}_{\Pi,\mathcal{A}}(\lambda) \;=\; \big|\Pr[b' = b^*] - 1/(\ell+1)\big|$.

Definition 4 (Content Deniability). *A DEEH scheme* Π *is* $\varepsilon_{\mathrm{CD}}$-content-deniable *if* $\mathsf{Adv}^{\mathrm{CD}}_{\Pi,\mathcal{A}}(\lambda) \le \varepsilon_{\mathrm{CD}}(\lambda)$ *for every PPT adversary* $\mathcal{A}$*. We say* Π *is* content-deniable *if* $\varepsilon_{\mathrm{CD}}(\lambda) = \mathit{negl}(\lambda)$.

A complementary (t,q)*-framework.* Definition 3 fixes the decoy pool in advance to size $\ell+1$. It is useful to complement it with an open-world notion that captures the adversary's residual uncertainty after coercion. We call Π (t,q)*-deniable* if every adversary with time complexity $\le t(\lambda)$ and $\le q$ secret-key queries to the honest party cannot rule out that a further, yet-undiscovered message is embedded in the challenger's ciphertext C. Trivially $q \le t$; the interesting regime is $q < \ell$, in which the adversary has consumed its query budget before exhausting the decoy pool. The two notions are related but not equivalent: CD security asks whether the adversary can *identify* the true message among a known pool, whereas (t,q)-deniability asks whether the adversary can *rule out further* hidden messages.

Example 2 (Canetti et al.). The sender-deniable PKE of [12] admits exactly one alternative opening. An adversary issuing a single key query receives the unique decoy opening, after which no further consistent message can be plausibly exhibited by the sender. Hence the scheme is $(\mathrm{poly}(\lambda), 1)$-deniable but not $(\mathrm{poly}(\lambda), q)$-deniable for any $q \ge 2$.

Example 3 (FBE). FBE [1] embeds $\ell+1$ messages via an underdetermined linear system over $\mathbb{F}_p$. An adversary holding $q \le \ell$ keys recovers q messages; the remaining $\ell+1-q$ slots are information-theoretically consistent with arbitrary field elements. FBE is therefore (∞, ℓ)-deniable. Residual uncertainty degrades to zero only once the adversary obtains all $\ell+1$ keys.

For PhantomCrypt specifically (Section 5), the set of candidate decoy openings is a subset of the $\binom{L}{k-1}$ choices of $k-1$ share positions among the L words of the cover text τ (the abscissae and the evaluation point being additionally hidden by the seed). The size of this candidate space is therefore a function of *both* the cover length L and the threshold k, and the (t,q) guarantee is meaningful only when it is super-polynomial in the security parameter. Two regimes must be distinguished, and conflating them is a real pitfall.

- *Constant threshold* ($k = O(1)$, the regime of our default instantiation). Here $\binom{L}{k-1} = \Theta(L^{k-1})$ grows only polynomially in L. With a polynomial cover

length $L = \text{poly}(\lambda)$, the candidate space is polynomial, so the open-world (t, q)-deniability claim does *not* hold: a polynomial-time adversary can, in principle, enumerate the position tuples. Constant k still yields the closed-world CD guarantee of Definition 3 (Theorem 4), which is about identifying the true message among a *fixed* pool, not about ruling out further hidden ones. We use constant k wherever only closed-world CD is required.

- *Growing threshold* ($k = \Theta(L)$, equivalently $k = \Theta(\lambda)$ for $L = \Theta(\lambda)$). Here $\binom{L}{k-1} = 2^{\Theta(L)}$ is super-polynomial, no polynomial-time adversary exhausts it, and PhantomCrypt is $(\text{poly}(\lambda), \text{poly}(\lambda))$-deniable in the open-world sense. This is the regime in which the residual-uncertainty argument is sound.

The choice is a genuine trade-off rather than a free parameter: the open-world guarantee costs a threshold (and hence an interpolation cost and a share-derivation cost) linear in the cover length, which we account for honestly in the evaluation (Section 8) rather than benchmarking only the cheaper constant-k setting.

3.3 Existence Deniability (ED)

Existence deniability captures the property that a transmitted document D is, in parts, indistinguishable from traffic produced by an innocent conventional (i.e., not deniable) cryptographic protocol, or transmission of randomly looking data (like an authentication token in a QR code). To this end, we introduce the concept of a *reference distribution* that captures the distribution of documents.

Definition 5 (Reference Distribution). *Fix a standard hybrid public-key encryption scheme $\Pi_{\text{ref}} = (\mathsf{KEM}, \mathsf{AEAD})$; concretely we take KEM = ML-KEM-768 and AEAD = AES-256-GCM. For a cover string τ and KEM public parameters PP, the* reference distribution $\mathcal{D}_{\text{ref}}(\tau, \mathsf{PP})$ *is the distribution of the tuple*

$$\big(\tau,\ C_{\text{kem}},\ \mathsf{IV}_1, \mathsf{IV}_2,\ \mathsf{AEAD.Enc}(K, r_1, \mathsf{IV}_1),\ \mathsf{AEAD.Enc}(K, r_2, \mathsf{IV}_2)\big),$$

where $(C_{\text{kem}}, K) \leftarrow \mathsf{KEM.Encaps}(\mathsf{PP})$, the IV_j are sampled as the scheme prescribes, and r_1, r_2 are uniform over the payload spaces of the corresponding PhantomCrypt payloads. In words, $\mathcal{D}_{\text{ref}}$ is exactly the wire output of Π_{ref} encrypting uniform payloads of the matching lengths, with the unmodified cover string prepended.

The reference distribution is thus a concrete, efficiently samplable object tied to a deployed protocol, not an abstract placeholder. The ED question is whether a PhantomCrypt ciphertext is distinguishable from a single draw of $\mathcal{D}_{\text{ref}}(\tau|\mathsf{PP})$: an attacker must decide whether each AEAD payload carried real content (the IE share s_{new} and the FBE container) or independent uniform strings of the matching length. Definition 7 formalizes this as a chosen-message indistinguishability game: the adversary picks the cover, and the challenge message, and the challenger flips a bit deciding whether to encrypt the real content or length-matched uniform payloads.

Remark 1 (Why two AEAD payloads). We present the wire format with two AEAD payloads C_1 wrapping the IE share s_{new} and C_2 wrapping the FBE container c_{FBE} purely for expository clarity, so that the existence-deniable component (IE) and the content-deniable component (FBE) appear as separate, independently analyzable frames. The two are not fundamental: a single AEAD payload that concatenates $s_{\text{new}}|c_{\mathsf{FBE}}$ under an unambiguous framing is equivalent for ED, since the reduction (Theorem 3) treats the payload only as a length-matched plaintext. The single-payload variant is, in fact, preferable in deployment, because the two-payload format makes the ciphertext length grow visibly with the decoy count ℓ (a mild ED signal, see Section 8); keeping the FBE container in one framed payload, or off the wire entirely as discussed in Section 5, removes that signal. We retain the two-payload presentation in the definitions only because it makes the layer-by-layer reduction easier to read.

Remark 2 (What ED does and does not cover). ED as defined is indistinguishability of the *ciphertext object* from $\mathcal{D}_{\text{ref}}$. It deliberately does not model traffic shape, timing, packet metadata, deployment context, or detection heuristics applied to a live channel; those depend on how the object is transported and are out of scope. A scheme can be ED-secure in our sense and still be flagged by a traffic-analysis system that keys on, e.g., message frequency. Two further caveats are essential to read the guarantee correctly. First, $\mathcal{D}_{\text{ref}}$ is the wire output of a *chosen* reference protocol, and ED says PhantomCrypt is indistinguishable from *that* object, not that the object is itself innocuous; if the reference (e.g., a natural-language cover string accompanied by a hybrid-encrypted attachment) is anomalous in a given deployment, ED does not repair it. Second, because the same cover string τ appears in both branches of the ED game (Definition 7), the game holds the plausibility of the cover string *fixed*: it certifies that PhantomCrypt adds no detectable signal beyond τ, not that τ looks natural. The latter, a quantitative metric of cover plausibility, is an explicit assumption here and an open problem (Section 9). We make these boundaries explicit because the motivation (Section 1) concerns real channels, whereas the guarantee is about the object placed on them; closing the gap requires a transport-layer and cover-modelling treatment we do not attempt.

Definition 6 (Admissible cover). *Fix the security parameter λ, the threshold k, and the random oracle $H : \{0,1\}^* \to \mathbb{F}_p$. A cover string $\tau = (w_1, \ldots, w_L)$ is* admissible *for PhantomCrypt if it satisfies all of:*

1. (Length.) *$L \geq \kappa(\lambda)$ for a public polynomial κ, with $L \geq k + 1$ so that the PRNG can select $k - 1$ distinct share positions and a fresh evaluation point can be derived.*
2. (Non-degeneracy.) *τ contains at least k* distinct *words. This excludes covers such as a single repeated token, which would fix every share $s_j = H(w_{i_j})$ to one value and collapse the share entropy even when the abscissae remain hidden.*
3. (Share separation.) *the induced share values $\{H(w_i)\}_{i=1}^{L}$ contain at least k distinct nonzero elements of $\mathbb{F}_p$. Since H is a random oracle and p is super-*

polynomial, a uniformly chosen τ of distinct words fails this only with probability $O(L^2/p) = negl(\lambda)$ (a birthday bound on H-collisions); admissibility makes the event explicit rather than assumed away.

4. (Frequency non-degeneracy.) *Write* $f_{\max}(\tau) = \max_{v \in \mathbb{F}_p} \frac{1}{L}\big|\{i : H(w_i) = v\}\big|$ *for the largest fraction of positions carrying a single share value, and* $H_\infty(\tau) = -\log_2 f_{\max}(\tau)$ *for the induced* share min-entropy *of one seed-driven selection. We require*

$$(k-1) \cdot H_\infty(\tau) = \omega(\log \lambda), \qquad \textit{equivalently} \qquad f_{\max}(\tau)^{k-1} = negl(\lambda).$$

Conditions 1–3 alone are not enough, which is why condition 4 is stated separately. A cover may contain k distinct words, and so pass conditions 2–3, while being overwhelmingly dominated by repetitions of a single token: boilerplate consisting of one word repeated $L - k$ times plus k distinct words once each is admissible under 1–3 but has $f_{\max} \approx 1$. Under such a cover the seed-driven selection concentrates on one share value with probability close to 1, the interpolation polynomial degenerates, and one particular candidate message acquires an anomalously large anonymity set. Lemma 4 quantifies exactly how much this costs and shows that condition 4 makes the cost negligible; Section 8 measures $H_\infty(\tau)$ on real text. Condition 4 is a restriction on word frequency*, not merely on word* distinctness*, which is the gap conditions 2–3 leave open.*

Both the honest party and the adversary can check admissibility from τ alone. All security statements quantify over admissible covers; an inadmissible cover trivially leaks the mechanism through structural impossibility and is excluded from the games below (the honest party, who chooses τ, is responsible for admissibility, as discussed in Section 3.4).

Definition 7 (ED Security Game). *For a DEEH scheme Π and adversary $\mathcal{A}$, the game* $\mathsf{Game}^{\mathrm{ED}}_{\Pi,\mathcal{A}}(\lambda)$ *proceeds as follows:*

1. *$\mathcal{A}$ submits an* admissible *cover string τ (Definition 6) and an arbitrary challenge message $m^* \in \mathcal{M}$ of its choosing. (Allowing $\mathcal{A}$ to choose m^* only strengthens the guarantee: the worst-case message cannot help the distinguisher.)*
2. *The challenger runs* $(\mathsf{PP}, sk_{\mathrm{den}}) \leftarrow \mathsf{Setup}(\lambda, \tau, \ell)$ *and samples* $b \xleftarrow{\$} \{0,1\}$.
3. *If $b = 1$: the challenger runs* $(C, dk_0) \leftarrow \mathsf{Enc}(\mathsf{PP}, sk_{\mathrm{den}}, m^*, \tau)$ *followed by* $\mathsf{AddDecoy}$ *on the remaining messages, embedding the real share s_{new} and the FBE container.*
   If $b = 0$: the challenger samples $C \leftarrow \mathcal{D}_{\mathrm{ref}}(\tau, \mathsf{PP})$*, i.e. the same hybrid wrapper over* length-matched uniform *payloads in place of s_{new} and c_{FBE}.*
   In either case the challenger returns (τ, C) to $\mathcal{A}$.
4. *$\mathcal{A}$ outputs a guess $b' \in \{0,1\}$.*

Define $\mathsf{Adv}^{\mathrm{ED}}_{\Pi,\mathcal{A}}(\lambda) = |\Pr[b' = b] - 1/2|$*. The adversary must thus decide whether the AEAD payloads carry real content (s_{new}, c_{FBE}) or independent uniform strings of the same lengths; since the two branches differ only in the AEAD plaintexts and the payloads are length-matched, the distinction reduces to AEAD indistinguishability under a KEM-derived key (Theorem 3).*

Definition 8 (Existence Deniability). *A DEEH scheme is $\varepsilon_{\mathrm{ED}}$-existence-deniable w.r.t. $\mathcal{D}_{\mathrm{ref}}$ if $\mathsf{Adv}^{\mathrm{ED}}_{\Pi,\mathcal{A}}(\lambda) \leq \varepsilon_{\mathrm{ED}}(\lambda)$ for every PPT $\mathcal{A}$. We say Π is existence-deniable if $\varepsilon_{\mathrm{ED}}(\lambda) = \mathit{negl}(\lambda)$.*

Remark 3. $\mathcal{D}_{\mathrm{ref}}$ is an explicit parameter: different instantiations arise from different baselines (e.g., PQC-KEM hybrid encryption, TLS records). The cover string τ can be a natural-language text, an image, or any carrier. In both branches of the ED game, the adversary sees the same τ; the challenge is solely whether the ciphertext was produced by PhantomCrypt or by the reference protocol.

3.4 Second-Order Deniability

We treat second-order deniability as a *definitional decomposition* rather than a standalone security game. The two halves are attacked in different phases by adversaries with different powers: ED faces a passive network observer who never obtains keys, while CD faces a coercer who obtains all disclosed keys. A single monolithic game would have to grant or withhold key access in a way that collapses one of the two settings, so we define 2OD as the conjunction and analyse each half in its own game. Security is understood in a passive sense: the honest party (challenger) sets up the system, chooses the cover text, fixes the sensitive message m^* and the decoys $m_1, \ldots, m_\ell$. This rules out trivial wins such as prescribing a cover text too short to carry any secret, or one that is a single repeated word (which would fix every share s_i, leaking information even though the x-values may remain hidden).

Definition 9 (Second-Order Deniability). *A DEEH scheme Π is* second-order deniable *with respect to $\mathcal{D}_{\mathrm{ref}}$ if it is both $\varepsilon_{\mathrm{CD}}$-content-deniable (Definition 4) and $\varepsilon_{\mathrm{ED}}$-existence-deniable (Definition 8) with $\varepsilon_{\mathrm{CD}}, \varepsilon_{\mathrm{ED}} = \mathit{negl}(\lambda)$. We adopt this conjunction as the working definition of 2OD, and we do not claim that it arises from a single monolithic game with its own reduction. Section 3.5 then strengthens it into a genuine indistinguishability notion at the level of the coercion disclosure, and shows that both components are necessary for that stronger notion, so the pairing is load-bearing rather than decorative.*

Proposition 1 (Characterization). *By Definition 9, a DEEH scheme is second-order deniable iff it is both content-deniable and existence-deniable. The intuition behind the decomposition: failure of ED lets a passive observer detect that a non-standard mechanism was used; failure of CD lets a coercer holding all disclosed keys identify the true message. Denying both is exactly denying that the mechanism was meaningfully used. This is a statement at the level of the definition, not a reduction.*

Remark 4 (Limits of the characterization). The decomposition captures the network observer and the coercer but, as Section 1.1 makes explicit, not a forensic adversary inspecting the device that runs the software. Against that adversary 2OD does not hold; CD does. We state this rather than fold it into the definition.

Lemma 1 (Two-stage composition). *Let Π be ε_1-ED-secure and ε_2-CD-secure. A two-stage adversary $\mathcal{A}$ that first observes (τ, C) and outputs a detection bit d, then (if $d = 1$) receives $\{dk_i\}_{i=0}^{\ell}$ and outputs b'_c, satisfies:*

$$\Pr[\textit{detect} \wedge \textit{identify}] \leq \left(\tfrac{1}{2} + \varepsilon_1\right)\left(\tfrac{1}{\ell+1} + \varepsilon_2\right).$$

If $\varepsilon_1, \varepsilon_2 = \textit{negl}(\lambda)$, this is at most $\frac{1}{2(\ell+1)} + \textit{negl}(\lambda)$: correctly detecting deniable encryption and *identifying the true message is only negligibly better than guessing among $2(\ell+1)$ possibilities.*

Proof. Condition on the event that the challenge ciphertext is generated by Π (i.e., $b = 1$ in the ED game). In the detection phase, $\mathcal{A}$'s view is the ED game, so for an adversary applying the ED-optimal distinguisher, $\Pr[d = 1 \mid b = 1] \leq 1/2 + \varepsilon_1$. Conditioned on $d = 1$, $\mathcal{A}$'s additional input is exactly the CD challenge of Definition 3, so $\Pr[b'_c = b_c \mid d = 1,\ b = 1] \leq 1/(\ell+1) + \varepsilon_2$. Chaining:

$$\begin{aligned}\Pr[\text{detect} \wedge \text{identify} \mid b = 1] &= \Pr[d{=}1 \mid b{=}1] \cdot \Pr[b'_c{=}b_c \mid d{=}1,\, b{=}1] \\ &\leq \left(\tfrac{1}{2} + \varepsilon_1\right)\left(\tfrac{1}{\ell+1} + \varepsilon_2\right).\end{aligned}$$

Expanding and dropping products of negligibles gives $\frac{1}{2(\ell+1)} + \text{negl}(\lambda)$. □

3.5 Making Content Deniability Load-Bearing

Definition 9 is a conjunction, and a conjunction invites the objection that one conjunct might be doing all the work. The objection has teeth. Any notion in which the coercion material is released only after an adversary has already accused a genuine object lets existence deniability cap the rate at which that gate opens, so the coercion stage is reached only with probability $2\varepsilon_{\text{ED}}$ and content deniability is never asked to do anything. A notion in which one component carries the whole burden is a notion in which the composition is decoration.

The remedy is to drop the gate and compare not two ciphertexts but two *complete coercion stories*, asking that the story a PhantomCrypt user tells under coercion be one that a user who never touched a deniability mechanism could have told. This is the level at which the conformance requirement flagged in Section 1.1 becomes statable, and the definitions below make it precise. We formalise the notion and prove both components necessary and (with one strengthening) sufficient; we do *not* prove PhantomCrypt conformant against a deployed container format, which remains the principal open problem of Section 9.

Definition 10 (Coercion view). *Let Π be a DEEH scheme with decoy budget ℓ, let τ be an admissible cover string, let $\boldsymbol{m} = (m_0, \ldots, m_\ell)$ be a vector of $\ell + 1$ pairwise distinct messages, and let $b^* \in \{0, \ldots, \ell\}$. The* coercion view *is*

$$\mathsf{View}_{\Pi}(\tau, \boldsymbol{m}, b^*) \;=\; \big(\mathsf{PP},\ \tau,\ C,\ dk_0, \ldots, dk_\ell\big),$$

obtained by running $\mathsf{Setup}(\lambda, \tau, \ell)$, then Enc on m_{b^*}, then $\mathsf{AddDecoy}$ on every m_i with $i \neq b^*$. This is precisely what a coercer walks away with, and it is the view

handed to the adversary in the CD game of Definition 3. Concretely, at $\ell = 4$ it is a tuple of public parameters, the cover article, the transmitted object, and five keys, each bundling the on-device KEM key with one FBE key; running Dec *five times yields five messages.*

Definition 11 (Innocent multi-opening reference). *By a* multi-record encrypted container *we mean the everyday object a password manager or an encrypted notebook produces: one file storing several independent records, each under its own key, with no record privileged over any other. A user of such a tool has nothing to hide by design, which is why it is the right innocent baseline. Formally, let Π_{mr} be such a container with $\ell+1$ records: a single transcript C_{mr} carries $m_0, \ldots, m_\ell$, and record key dk_i opens C_{mr} to m_i. The container has no notion of a true record, so its $\ell + 1$ openings are symmetric by construction, meaning that permuting which record is called first leaves the joint distribution of transcript and keys unchanged. Write*

$$\mathsf{View}_{\mathrm{mr}}(\tau, \boldsymbol{m}) \;=\; \big(\mathsf{PP},\ \tau,\ C_{\mathrm{mr}},\ dk_0, \ldots, dk_\ell\big)$$

for the owner's own view of such a container, all of whose record keys have been seized. The ciphertext marginal *of Π_{mr} is the distribution of the transcript component C_{mr} alone, obtained by discarding the public parameters, the cover string, and all keys.*

Definition 12 (Strong second-order deniability). *A DEEH scheme Π is* strongly second-order deniable *relative to Π_{mr} if for every admissible cover τ and every tuple $\boldsymbol{m}$ of $\ell + 1$ pairwise distinct messages the joint ensembles*

$$\big\{(\mathsf{View}_{\Pi}(\tau, \boldsymbol{m}, b^*),\ b^*) \;:\; b^* \xleftarrow{\$} \{0, \ldots, \ell\}\big\} \quad \textit{and} \quad \big\{(\mathsf{View}_{\mathrm{mr}}(\tau, \boldsymbol{m}),\ \hat{b}) \;:\; \hat{b} \xleftarrow{\$} \{0, \ldots, \ell\}\big\}$$

are computationally indistinguishable, with advantage $\kappa(\lambda) = \mathit{negl}(\lambda)$. In the second ensemble the index $\hat{b}$ is sampled independently of the view.

The independence of $\hat{b}$ in the ideal ensemble is the whole design. It forces the *pair*, not just the view, to be simulatable, and so it says two things at once: the coerced party's entire story is one an innocent container owner could have told, and no opening inside that story is marked as the true one. The first is existence deniability at the level of the disclosure rather than the wire; the second is content deniability. Two failure modes make the point concrete. Suppose the FBE keys were emitted in encryption order, so dk_0 is always the true one. The view is still a perfectly ordinary-looking container, so the story conforms, but the pair $(\mathsf{View}_{\Pi}, b^*)$ is trivially distinguishable from $(\mathsf{View}_{\Pi}, \hat{b})$: read the index off the key order. Now suppose instead the keys are symmetric, so no key carries a field, a length, or an ordering identifying it as the one Enc emitted, but the ciphertext is a bare $\mathbb{F}_p$ vector. The pair is now index-hiding, but no innocent container has that shape, so conformance fails. Definition 9 sees only the second failure; the next theorem shows Definition 12 sees both.

Theorem 1 (Both components are necessary). *Let Π be strongly second-order deniable relative to Π_{mr} with distinguishing advantage at most $\kappa(\lambda)$. Then*

1. *Π is content-deniable, with $\mathsf{Adv}^{\mathrm{CD}}_{\Pi,\mathcal{A}}(\lambda) \le \kappa(\lambda)$ for every PPT $\mathcal{A}$, and*
2. *Π is existence-deniable with respect to the ciphertext marginal of Π_{mr}, with $\mathsf{Adv}^{\mathrm{ED}}_{\Pi,\mathcal{A}}(\lambda) \le \kappa(\lambda)$.*

Proof. (1) Let $\mathcal{A}$ be a CD adversary. Build a distinguisher $\mathcal{D}$ that, on a pair (V, i), runs $\mathcal{A}(V)$ and outputs 1 exactly when $\mathcal{A}$'s guess equals i. On the real ensemble the pair is $(\mathsf{View}_{\Pi}(\tau, \boldsymbol{m}, b^*), b^*)$, so $\Pr[\mathcal{D} = 1] = \Pr[\mathcal{A} \text{ wins } \mathsf{Game}^{\mathrm{CD}}]$. On the ideal ensemble $\hat{b}$ is uniform and independent of $\mathsf{View}_{\mathrm{mr}}$, so $\mathcal{A}$'s guess hits it with probability exactly $1/(\ell+1)$, whatever $\mathcal{A}$ does. Strong second-order deniability bounds the difference of the two acceptance probabilities by κ, and that difference is $\left|\Pr[\mathcal{A} \text{ wins}] - 1/(\ell+1)\right| = \mathsf{Adv}^{\mathrm{CD}}_{\Pi,\mathcal{A}}$.

(2) Let $\mathcal{A}$ be an ED adversary. Build $\mathcal{D}$ that discards the decoy keys and the index, runs $\mathcal{A}$ on (PP, τ, C), and echoes its output. The real projection of the pair is the $b = 1$ branch of the ED game and the ideal projection is the ciphertext marginal of Π_{mr}, so $\mathcal{D}$'s advantage is exactly $\mathcal{A}$'s. □

Sufficiency needs one ingredient more, and naming it is the point of the exercise. Content deniability as defined is an *unpredictability* statement: no adversary guesses b^* better than blindly. Definition 12 asks for an *indistinguishability* statement about the pair, which is strictly stronger, since a view could leak partial information about b^*, say its parity, without ever letting the index be pinned down. The following definition isolates the stronger property, and the one after it isolates the structural half.

Definition 13 (Index hiding). *Π is* ε-index-hiding *if for every admissible τ and every $\boldsymbol{m}$, the ensembles $(\mathsf{View}_{\Pi}(\tau, \boldsymbol{m}, b^*), b^*)$ and $(\mathsf{View}_{\Pi}(\tau, \boldsymbol{m}, b^*), \hat{b})$, with $b^*, \hat{b}$ independent and uniform on $\{0, \ldots, \ell\}$, are indistinguishable with advantage at most $\varepsilon(\lambda)$.*

Definition 14 (Conformance). *Π is* κ-conformant *with Π_{mr} if for every admissible τ and every $\boldsymbol{m}$, the distribution of $\mathsf{View}_{\Pi}(\tau, \boldsymbol{m}, b^*)$ with b^* uniform is indistinguishable from $\mathsf{View}_{\mathrm{mr}}(\tau, \boldsymbol{m})$ with advantage at most $\kappa(\lambda)$.*

Theorem 2 (Index hiding and conformance suffice). *If Π is ε-index-hiding and κ-conformant with Π_{mr}, then Π is strongly second-order deniable relative to Π_{mr} with advantage at most $\varepsilon(\lambda) + \kappa(\lambda)$.*

Proof. Three hybrids. $\mathsf{H}_0 = (\mathsf{View}_{\Pi}(\tau, \boldsymbol{m}, b^*), b^*)$ is the real ensemble. $\mathsf{H}_1 = (\mathsf{View}_{\Pi}(\tau, \boldsymbol{m}, b^*), \hat{b})$ replaces the true index by an independent uniform one, and $|\mathsf{H}_0 - \mathsf{H}_1| \le \varepsilon$ by index hiding. $\mathsf{H}_2 = (\mathsf{View}_{\mathrm{mr}}(\tau, \boldsymbol{m}), \hat{b})$ is the ideal ensemble. Since $\hat{b}$ is independent of the view in both H_1 and H_2, a distinguisher can sample it itself, so $|\mathsf{H}_1 - \mathsf{H}_2| \le \kappa$ by conformance. The triangle inequality gives the bound. □

Remark 5 (What PhantomCrypt does and does not establish here). Index hiding is supplied by the algebra, and with an explicit constant. The proof of Theorem 4 does not merely bound a guessing probability: it bounds a *statistical*

distance, showing that s_{new} is within $\delta(\lambda)$ of uniform and, up to $2\delta(\lambda)$, independent of which message sits at $P(0)$ (Lemma 3), and that the FBE container is invariant under permutation of its openings (Lemma 5). Those are exactly the ingredients Definition 13 asks for, so PhantomCrypt is ε-index-hiding with $\varepsilon = 2\delta(\lambda) + 2^{-(k-1)H_\infty(\tau)} + \varepsilon_{\text{FBE}}(\lambda)$. Conformance is a different matter. It asks that the disclosed tuple be indistinguishable from the view of a real multi-record container, which is a claim about a *deployed format*, not about Π alone, and it cannot be discharged without fixing that format. We therefore state conformance as a definition and a target rather than a theorem, and record it as the principal open problem. This is the formal content of the informal conformance requirement raised in Section 1.1.

3.6 Independence of CD and ED

Lemma 2 (Independence). *Content deniability and existence deniability are logically independent.*

Proof. CD $\not\Rightarrow$ ED. FBE [1] embeds multiple messages in a single ciphertext via underdetermined linear systems, achieving information-theoretic CD. However, its ciphertext is a raw vector of field elements, trivially distinguishable from standard protocol traffic: $\mathsf{Adv}^{\text{ED}} = 1/2 - \text{negl}(\lambda)$.

ED $\not\Rightarrow$ CD. IE [3] embeds a secret message as the constant term of a polynomial interpolated through shares derived from an unmodified cover text, with the transmitted share carried in a standard hybrid encryption envelope. IE achieves ED but embeds only a single message with no prepared decoy keys. In the CD game with $\ell \geq 1$, the adversary identifies m^* with probability 1. □

Remark 6 (IE does not provide CD). In IE, a party could construct a different polynomial through the same shares, yielding a different constant term. However, the CD game gives the adversary all $\ell + 1$ decoy keys. IE has no decoy keys: an ad hoc alternative does not constitute a decoy key in Definition 1. Read as a DEEH scheme, IE has one payload and no non-trivial $\mathsf{AddDecoy}$, so its decoy budget is $\ell = 0$. At $\ell = 0$ the CD game presents a single message, the guessing baseline is 1, and $\mathsf{Adv}^{\text{CD}} = 0$ vacuously: IE supplies existence deniability and nothing on which content deniability could bite. That is the gap the composition with FBE closes, and it is why a decoy budget $\ell \geq 1$ is part of the syntax rather than an afterthought.

4 Related Work

We place related work after the formal definitions so that existing schemes can be classified by which of the CD and ED they achieve.

Content deniability schemes. Canetti et al. [12] introduced deniable encryption; subsequent work [11] achieved full bi-deniability via iO. An and Zhang [6] introduced dual-deniable PKE; An et al. [5] achieved leakage-resilient deniable encryption. FBE [1] supports arbitrary decoys via underdetermined linear systems,

achieving information-theoretic CD. Honey Encryption [30] renders incorrect-key decryptions plausible. Hidden volumes (VeraCrypt [38], Shufflecake [8]) provide CD at the storage level. Canetti et al. [10] introduced deniable secret sharing, exploiting the same algebraic structure as FBE. All produce ciphertexts distinguishable from standard traffic, failing ED.

Existence deniability and steganography. Traditional steganography [16, 29, 39] and linguistic steganography [17, 18] embed hidden messages in cover media. Provably secure generative steganography: Kaptchuk et al. [31] via GPT-2; De Witt et al. [41] via minimum entropy coupling; Ding et al. [22] via distribution copies. IE [3] derives shares from an unmodified cover text within a hybrid encryption envelope. These achieve ED but not CD. IE is also not post-quantum on its own: nothing in the construction fixes the carrier, but the reported instantiation uses RSA-based hybrid encryption, which Shor's algorithm breaks, so the migration is left open there and the post-quantum treatment belongs to the composition presented here.

Anamorphic encryption. Persiano et al. [34] introduced anamorphic encryption. Banfi et al. [9] proposed robust anamorphic encryption; Catalano et al. [15] extended it to hybrid encryption, IBE, and FHE; Catalano et al. [14] proved generic impossibility for stateless black-box anamorphic encryption with large message spaces; Persiano et al. [35] introduced public-key anamorphism; Choi et al. [19] unified receiver notions; Deo and Libert [21] gave the first lattice-based construction; Carnemolla et al. [13] and Dodis and Goldin [23] constructed anamorphic-resistant PKE. Anamorphic encryption is the closest competitor and deserves a careful contrast.

Its central strength is precisely the axis on which PhantomCrypt is weakest. An anamorphic ciphertext is produced by, and is a valid ciphertext of, a *standard* PKE scheme; the only extra secret is a double key. A forensic adversary inspecting a seized device, therefore, finds ordinary, widely deployed encryption software, so the Stage-2 collapse we describe in Section 1.1 does not arise in the same way: there is no non-standard tool to discover. The discoverability of deniable tooling on seized devices is a documented, practical attack surface [28], which is exactly why this axis matters. PhantomCrypt does not match this. Where PhantomCrypt is stronger is in capacity and symmetry under full-key disclosure. Anamorphic encryption embeds a single covert message alongside the normal one, and normal and covert decryption are structurally asymmetric, so an adversary who obtains all keys can, in general, identify which mode is "normal." The impossibility results of [14] further bound the covert capacity of stateless black-box constructions. PhantomCrypt instead provides $\ell + 1$ structurally symmetric openings (Theorem 4) with no distinguished "normal" mode and super-polynomial residual uncertainty (Section 3.2), which the two-message anamorphic setting cannot offer. The two designs thus trade off along orthogonal axes: anamorphic encryption buys software-forensic deniability at the cost of capacity and symmetry, PhantomCrypt the reverse.

Table 2: Anamorphic encryption versus PhantomCrypt along the axes where they differ. The two are complementary rather than ordered: each is strong exactly where the other is weak.

	Anamorphic	**PhantomCrypt**
Covert capacity	one covert message alongside the normal one	$\ell + 1$ messages in one container
Openings under full-key disclosure	asymmetric: normal vs. covert mode distinguishable	$\ell + 1$ structurally symmetric openings; none marked true
Residual uncertainty after disclosure	bounded by the two-message structure [14]	super-polynomial in the growing-k regime (Section 3.2)
Software-forensic deniability (device)	yes: ciphertext of a standard, widely deployed 2OD PKE; no non-standard tool to find	no: a non-standard tool on the device collapses at Stage 2 (Section 1.1)

Deniability in secure messaging. A parallel line of work studies deniability for interactive secure messaging rather than for stored or transmitted ciphertexts. Unger and Goldberg [37] decompose it into *message repudiation* (a party can deny sending a given message) and *participation repudiation* (a party can deny taking part in the conversation at all), and note that the two are orthogonal: a protocol may achieve either, both, or neither. The decomposition mirrors ours, with message repudiation aligned to CD and participation repudiation to ED; the secure-messaging SoK [36] surveys the surrounding notions. The mechanism and adversary differ, however. There, deniability means a conversation transcript can be forged by anyone holding only public keys, so that a rational judge shown the transcript obtains no transferable cryptographic proof, and the adversary evaluates after-the-fact evidence in an authenticated key exchange. Our CD instead faces a coercer who already holds all disclosed keys and must be served a plausible fake plaintext, while our ED faces a passive network observer who must fail to distinguish the ciphertext object from a fixed reference protocol. Recent work formalizes finer-grained messaging-deniability models, including distinguishers that know every secret key (the "big brother" setting) [26]; these are stronger adversaries than our ED targets, and we do not subsume them. What this lineage does not address is the feasibility question at the centre of our analysis: its repudiation notions never require a chosen fake to be openable under some key, so they never meet the seed-entropy threshold of Section 7.

Obfuscated channels and censorship resistance. Protocol obfuscation (obfs4 [7], ScrambleSuit [40]) disguises Tor traffic. Fenske and Johnson [24, 25] formalized Fully Encrypted Protocols; Günther et al. [27] formalized obfuscated key exchange with ML-KEM. Wu et al. [42] documented Great Firewall detection of

fully encrypted traffic. These are network-layer ED mechanisms that do not address CD.

At-compromise and socio-technical security. Albrecht et al. [4] introduced alert blindness from ethnographic fieldwork during Kenya's 2024 protests; Collins et al. [20] found that cryptographic deniability is never invoked by defendants in Swiss courts. Complementary socio-technical work argues that the *credibility* of a denial, not its cryptographic soundness, often determines outcomes under coercion [32]: a mathematically perfect decoy is worthless if the surrounding context makes it implausible. This sharpens our own scoping CD by providing cryptographic symmetry across openings, but the plausibility of the decoy *messages* (as opposed to the cover text) is a deployment responsibility we do not model, and constructing credible decoys is an open usability problem (Section 9).

Deniability via exchangeability in other modalities. Beyond encryption and storage, deniability can be obtained from structural *exchangeability*: making multiple candidate explanations symmetric so that none is distinguished. Recent work on plausibly deniable computation exploits exchangeability and multi-location symmetry to this end in a different modality [2]. PhantomCrypt's CD rests on the same principle in the algebraic setting, the $\ell + 1$ FBE openings are structurally symmetric (Lemma 5) so that no key is marked as "true", which suggests exchangeability is a recurring design pattern for content deniability across domains, even where the underlying mechanisms differ.

Position of PhantomCrypt. The schemes above each target CD or ED, but not both, with a single object and matching game-based definitions. Our contribution is primarily conceptual and definitional: the CD/ED decomposition, the proof that the two are logically independent, and the observation that they rest on a shared secret-sharing substrate over $\mathbb{F}_p$, which is what makes the composition natural rather than ad hoc. PhantomCrypt is the concrete witness that the two can be combined into a single ciphertext that is network-level indistinguishable from standard hybrid encryption while admitting $\ell + 1$ structurally symmetric openings. We do not claim a new cryptographic primitive; IE and FBE are prior work, and the value here lies is the framework, the independence result, and the demonstration that a single algebraic foundation supports both properties simultaneously.

5 PhantomCrypt Construction

PhantomCrypt instantiates the DEEH syntax (Definition 1) as follows: the cover string τ is a natural-language text $T = (w_1, \ldots, w_L)$; $\mathcal{M} = \mathbb{F}_p$; sk_{den} is the IE secret $\mathsf{SK}_{\text{ie}} = (x_0, k)$, pre-shared via IE [3] and never written to the device; each decoy key bundles the on-device decapsulation key with one FBE key, $dk_i = (sk_{\text{kem}}, \mathsf{SK}_{\mathsf{FBE},i})$; and $\mathcal{D}_{\text{ref}}$ is the output of a standard PQC-KEM hybrid encryption. Bundling sk_{kem} rather than naming it separately is what lets Dec

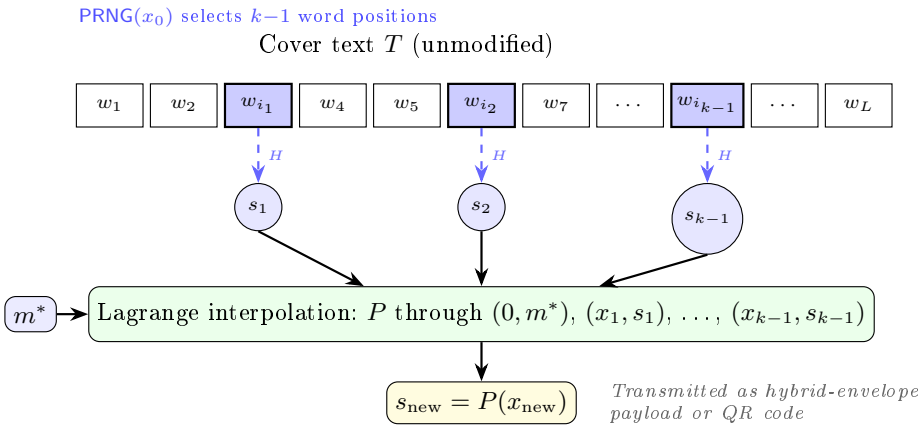

Fig. 2: IE share derivation. The seed drives two separate derivations: a PRNG seeded by x_0 selects $k-1$ words from cover text T, each hashed to an *ordinate* $s_j = H(w_{i_j})$, while the *abscissae* come from an independent hash chain $x_j = H(x_{j-1})$ on the same seed, continued past used values to a fresh x_{new}. A polynomial P is interpolated through the points (x_j, s_j) and $(0, m^*)$. That the two sources are independent given T is what Lemma 4 turns on: a repetition-heavy cover collapses the ordinates and leaves only one source. The transmitted share $s_{\text{new}} = P(x_{\text{new}})$ is pseudorandom and embeddable in any channel admitting random payloads. The cover text is never modified.

keep a single key slot, and it costs nothing: the same sk_{kem} simply repeats across all $\ell + 1$ decoy keys, and a coercer who seizes the device holds it anyway.

5.1 Building Blocks

Invisible Encryption (IE) [3]. Given a cover text $T = (w_1, \ldots, w_L)$ and a secret seed x_0, IE deterministically selects $k-1$ words from T using a PRNG seeded by x_0. Each selected word w_{i_j} produces a share $s_j = H(w_{i_j}) \in \mathbb{F}_p$ at abscissa x_j. A degree-$(k-1)$ polynomial P is interpolated through Equation (1), and the transmitted share $s_{\text{new}} = P(x_{\text{new}})$ is carried as the AEAD-wrapped payload within a standard hybrid encryption envelope. The cover text is *never modified*: shares are derived from it, not embedded into it. Figure 2 illustrates this process.

Provides: ED. The transmitted ciphertext (KEM ciphertext + AEAD payload) is syntactically and distributionally consistent with standard hybrid encryption. *Does not provide:* CD. The polynomial P is uniquely determined by x_0 and s_{new}. An adversary who obtains x_0 recovers m^* with certainty; there are no prepared decoy keys.

False-Bottom Encryption (FBE) [1]. FBE embeds multiple messages within a single ciphertext container $c_{\mathsf{FBE}} \in \mathbb{F}_p^N$ using underdetermined linear equations.

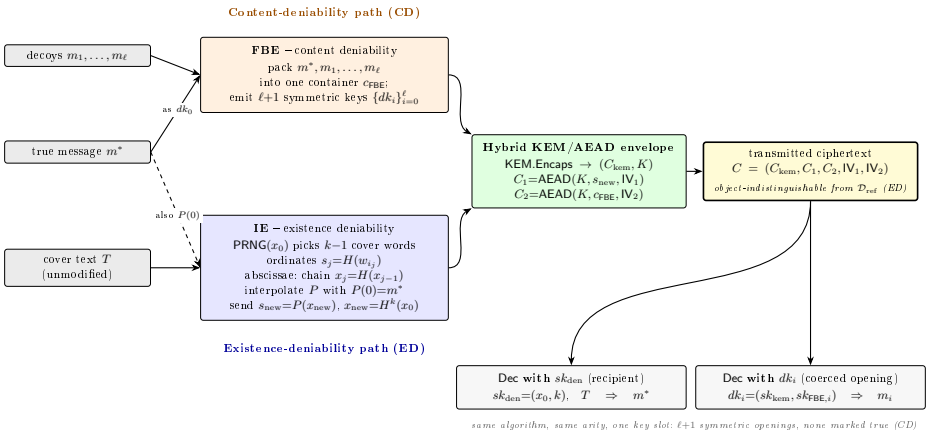

Fig. 3: PhantomCrypt process flow. FBE packs m^* and the ℓ decoys into one container with $\ell+1$ symmetric keys (CD); IE derives s_{new} from the unmodified cover text with m^* at $P(0)$ (ED). A hybrid KEM/AEAD envelope wraps both into C, object-indistinguishable from $\mathcal{D}_{\text{ref}}$. The true message is dual-embedded (IE channel via x_0; FBE container under dk_0), and both openings are the same call to Dec.

Each message m_i is associated with a decryption key $\mathsf{SK}_i = (I_{c,i}, I_{r,i}, \text{pos}_i)$ specifying which container elements and key-base elements to combine. The system of equations is underdetermined, ensuring that the container is consistent with all $\ell+1$ messages simultaneously. All keys are structurally symmetric: no key is distinguishable as the "true" one from the ciphertext and keys alone.

Provides: CD. Given c_{FBE} and all $\ell+1$ keys, an adversary cannot determine which message was designated as true. *Does not provide:* ED. The container consists of raw field elements with no resemblance to standard protocol traffic. Figure 3 shows the overall process flow.

5.2 Composition

FBE embeds m^* and ℓ decoys into c_{FBE} with symmetric keys (providing CD). IE independently derives shares from the unmodified cover text and computes s_{new} (providing ED). A PQC-KEM establishes a session key K_{session} that encrypts two AEAD payloads:

$$C_1 = \mathsf{AE.Enc}(K_{\text{session}}, s_{\text{new}}, \mathsf{IV}_1), \quad C_2 = \mathsf{AE.Enc}(K_{\text{session}}, c_{\mathsf{FBE}}, \mathsf{IV}_2).$$

The full ciphertext is $C = (C_{\text{kem}}, C_1, C_2, \mathsf{IV}_1, \mathsf{IV}_2)$.

True decryption requires $\mathsf{SK}_{\text{ie}} = (x_0, k)$, sk_{kem}, and the cover text T: the receiver decapsulates the KEM, decrypts C_1 to recover s_{new}, re-derives the $k-1$ shares from T using the pre-shared seed x_0, and reconstructs $m^* = P(0)$ via Lagrange interpolation.

Decoy decryption requires sk_{kem} and an FBE key $\mathsf{SK}_{\mathsf{FBE},i}$: the receiver decapsulates the KEM, decrypts C_2 to recover c_{FBE}, and applies FBE decryption. Since m^* is also embedded in FBE, decoy decryption with the FBE key for m^* also recovers m^* but the key is structurally indistinguishable from any other FBE key, providing no information about which message is true.

Remark 7 (PhantomCrypt coercion model: what the coercer recovers). In the CD game (Definition 3), the adversary receives all $\ell + 1$ decoy keys $dk_i = (sk_{\text{kem}}, \mathsf{SK}_{\mathsf{FBE},i})$ and runs Dec on each: every call decapsulates C_{kem}, decrypts C_2, and runs FBE decryption, so all $\ell + 1$ messages come out and every call is a call to the same algorithm with the same arity. Only $sk_{\text{den}} = \mathsf{SK}_{\text{ie}} = (x_0, k)$ is withheld. In particular the coercer recovers m^* itself it is one of the $\ell + 1$ openings so CD conceals the true message's *designation*, not its content. The design therefore assumes m^* is a plausible member of the decoy set, a condition on the credibility of the decoys rather than on the cryptography.

Why composition is natural. Information-theoretic CD requires algebraic structure admitting multiple consistent decryptions, while ED requires ciphertexts matching a standard protocol distribution. These goals are in tension: raw algebraic objects do not resemble protocol traffic, and standard ciphertexts typically admit unique decryptions. PhantomCrypt resolves this by composing FBE (algebraic structure for CD) with IE+KEM (protocol-conformant format for ED). Alternative architectures exist (e.g., a KEM with Canetti-style deniable encryption [12]), but PhantomCrypt achieves CD with negligible advantage in the random oracle model (statistical, given the seed x_0 is not disclosed and has entropy above $\log_2 p$; see Theorem 4). PhantomCrypt satisfies Definition 2 under the natural instantiation $sk_{\text{den}} = \mathsf{SK}_{\text{ie}}$ and $dk_i = (sk_{\text{kem}}, \mathsf{SK}_{\mathsf{FBE},i})$; the proof is in Appendix A.

6 Security Analysis

We establish PhantomCrypt's CD and ED properties via reductions to its building blocks.

6.1 Existence Deniability

Theorem 3 (ED Security). *Let $\mathcal{D}_{\text{ref}}(\tau)$ be the reference distribution of Definition 5 over an admissible cover τ. Assume the KEM is* IND-CCA *secure with advantage ε_{KEM}, the AEAD is* IND-CPA *secure with advantage $\varepsilon_{\text{AEAD}}$ under a uniformly random key, and AEAD nonces are sampled fresh and independently for each frame. Then for the two-payload wire format, $\varepsilon_{\text{ED}} \le \varepsilon_{\text{KEM}} + 2\varepsilon_{\text{AEAD}}$, the factor 2 arising from the two independently encrypted payloads via a hybrid over (C_1, C_2). (IND-CPA suffices because the ED game gives $\mathcal{A}$ no decryption oracle on the challenge envelope; IND-CCA is required of the KEM only to argue that the encapsulated key is replaceable by a uniform key in the presence of $\mathcal{A}$'s view, and may be relaxed to IND-CPA-KEM if no related-ciphertext queries are modeled.)*

Proof. Game 0: The real ED game ($b = 1$). The challenger runs PhantomCrypt: encapsulates to obtain $(C_{\text{kem}}, K_{\text{session}})$, and encrypts $C_1 = \mathsf{AE.Enc}(K_{\text{session}}, s_{\text{new}}, \mathsf{IV}_1)$ and $C_2 = \mathsf{AE.Enc}(K_{\text{session}}, c_{\mathsf{FBE}}, \mathsf{IV}_2)$ with fresh independent nonces $\mathsf{IV}_1, \mathsf{IV}_2$.

Game 1: Replace K_{session} with $\tilde{K} \xleftarrow{\$} \{0,1\}^{256}$. By IND-CCA of the KEM, $|\Pr[\text{G0}] - \Pr[\text{G1}]| \leq \varepsilon_{\text{KEM}}$.

Game 2: Replace the C_1 plaintext s_{new} with a uniform length-matched r_1. By IND-CPA of AEAD under the now-uniform $\tilde{K}$ and a fresh nonce, $|\Pr[\text{G1}] - \Pr[\text{G2}]| \leq \varepsilon_{\text{AEAD}}$.

Game 3: Replace the C_2 plaintext c_{FBE} with a uniform length-matched r_2. Since IV_2 is independent of IV_1, the second AEAD instance is keyed and nonced independently, so a second IND-CPA step gives $|\Pr[\text{G2}] - \Pr[\text{G3}]| \leq \varepsilon_{\text{AEAD}}$.

In Game 3, the ciphertext is a KEM ciphertext followed by two AEAD encryptions of uniform length-matched payloads under a uniform key with fresh nonces: exactly $\mathcal{D}_{\text{ref}}(\tau, \mathsf{PP})$ (the $b = 0$ branch). The cover text τ is identical in all games (IE never modifies it; $\mathcal{A}$'s submitted τ is used throughout), and length-matching of the payloads makes Games 2–3 valid IND-CPA reductions. By the triangle inequality, $\varepsilon_{\text{ED}} \leq \varepsilon_{\text{KEM}} + 2\varepsilon_{\text{AEAD}}$. The single-payload variant (Remark 1) replaces Games 2–3 with one IND-CPA step, giving $\varepsilon_{\text{ED}} \leq \varepsilon_{\text{KEM}} + \varepsilon_{\text{AEAD}}$. □

Remark 8 (Nonce model and misuse). The reduction assumes AEAD nonces are fresh and independent in both the construction and $\mathcal{D}_{\text{ref}}$, so that each frame is a fresh IND-CPA instance. Under GCM, nonce reuse across two frames under the same session key is catastrophic for indistinguishability (it leaks the XOR of the two keystreams and forges the authenticator), which would break ED regardless of the KEM. A deployment must therefore sample $\mathsf{IV}_1, \mathsf{IV}_2$ independently at random (or use a misuse-resistant AEAD); since $\mathcal{D}_{\text{ref}}$ samples nonces "as the scheme prescribes," the construction and the reference must use the identical nonce policy for the two distributions to coincide.

6.2 Content Deniability

Lemma 3 (Conditional uniformity of the IE share). *Model H and the seed-driven PRNG as random oracles. Fix a cover text T and threshold $k \geq 3$, and let the seed $x_0 \xleftarrow{\$} \{0,1\}^\lambda$ be uniform and unknown to the adversary. Condition on the adversary's full view in the CD game, which includes T and hence the entire set of candidate share values $\{H(w_i)\}_{i=1}^{L}$ (these are public and adversary-computable), but* not *x_0. Then for every fixed message $m \in \mathbb{F}_p$, the transmitted share $s_{\text{new}} = P(x_{\text{new}})$ is, conditioned on this view, statistically close to uniform over $\mathbb{F}_p$, with statistical distance at most $\delta(\lambda)$ where $\delta(\lambda) \to 0$ as the seed entropy exceeds $\log_2 p$ by a security margin (made precise in the proof). In particular, the conditional distribution of s_{new} is, up to $\delta(\lambda)$, independent of m.*

Proof. We first correctly identify the source of randomness. Writing L_j for the Lagrange basis polynomials on abscissae $\{0, x_1, \ldots, x_{k-1}\}$,

$$s_{\text{new}} \;=\; m \cdot L_0(x_{\text{new}}) \;+\; \sum_{j=1}^{k-1} s_j \cdot L_j(x_{\text{new}}), \qquad s_j = H(w_{i_j}).$$

Unlike in classical Shamir sharing, the values s_j are *not* hidden: the words $w_1, \ldots, w_L$ are the public cover text, so the adversary can compute every candidate $H(w_i)$. The hiding, therefore, cannot come from the s_j, and an argument that declares them "uniform and unknown" would be invalid in this model (this is exactly the computational, not information-theoretic, posture of IE [3]). The randomness that the adversary lacks is the seed x_0, which determines (i) which $k-1$ of the L words are selected as shares, (ii) the abscissae $x_1, \ldots, x_{k-1}$, and (iii) the evaluation point x_{new}, all via the random oracles.

Fix any candidate message m. Consider the map $\Phi_m : \{0,1\}^\lambda \to \mathbb{F}_p$ sending a seed x_0 to the share s_{new} it induces (with m placed at $P(0)$ and T fixed). In the random oracle model the selection indices and the abscissae are independent uniform draws, so Φ_m behaves as a random function from the seed space $\{0,1\}^\lambda$ to $\mathbb{F}_p$ (collision-avoidance for the k abscissae removes only a $\text{negl}(\lambda)$ fraction of seeds and contributes to δ). For a random function into $\mathbb{F}_p$ evaluated at a uniform input, the output is statistically close to uniform: the distribution of $\Phi_m(x_0)$ over uniform x_0 has statistical distance from uniform bounded by a quantity that vanishes once the seed space exceeds the field by a security margin, i.e. $\lambda \geq \log_2 p + s$ for a statistical parameter s, giving $\delta(\lambda) \leq 2^{-\Omega(s)}$. (This is the same S/p coverage law that governs how many seeds explain a fixed transcript as a given message; see the discussion following Theorem 4.) Since this bound is independent of which m was used m only shifts the affine term $m \cdot L_0(x_{\text{new}})$, and Φ_m is a random function for every fixed m the conditional distributions of s_{new} for any m, m' are each within $\delta(\lambda)$ of uniform and hence within $2\delta(\lambda)$ of each other. □

Theorem 4 (CD Security). *Model H and the PRNG as random oracles, let p have bit-length λ, and let the seed be drawn from $\{0,1\}^{\lambda+s}$, where $s = s(\lambda) \geq 1$ is a monotone non-decreasing* statistical security parameter*, so the seed space is s bits wider than the field. Suppose the FBE container is permutation-symmetric in the sense of Lemma 5 below. Then PhantomCrypt is ε_{CD}-content-deniable with*

$$\varepsilon_{\text{CD}}(\lambda) \;\leq\; 2\,\delta(\lambda) \;+\; 2^{-(k-1)H_\infty(\tau)} \;+\; \varepsilon_{\text{FBE}}(\lambda) \;=\; \mathit{negl}(\lambda),$$

where $\delta(\lambda) \leq 2^{-\Omega(s)}$ is the statistical-distance bound of Lemma 3, $2^{-(k-1)H_\infty(\tau)} = f_{\max}(\tau)^{k-1}$ is the cover-concentration term of Lemma 4, negligible by condition 4 of Definition 6, and ε_{FBE} is the FBE permutation-distinguishing advantage. The bound is negligible in λ if and only if *$s = \omega(\log \lambda)$, the cover satisfies condition 4 of Definition 6, and ε_{FBE} is itself negligible. It is* not *negligible for constant s, where $2^{-\Omega(s)}$ is a constant, nor for $s = \Theta(\log \lambda)$, where it is only inverse-polynomial. We state the equivalence rather than the one-sided claim because the*

construction itself forces no growth rate on s: a deployment is free to hold s constant, and thereby free to run a correct implementation with no asymptotic guarantee at all. A deployment setting $s = \lambda$, so that the seed is 2λ bits, obtains $\delta(\lambda) \le 2^{-\Omega(\lambda)}$. In the idealized limit $s \to \infty$ with unbounded lexical spread, H a perfect random oracle and FBE perfectly symmetric, $\varepsilon_{\mathrm{CD}} \to 0$.

Proof. In the CD game the adversary receives C and all $\ell + 1$ decoy keys $dk_i = (sk_{\mathrm{kem}}, \mathsf{SK}_{\mathsf{FBE},i})$, but not the deniable secret $sk_{\mathrm{den}} = \mathsf{SK}_{\mathrm{ie}} = (x_0, k)$. Using the sk_{kem} carried in any dk_i it decrypts C_1 to obtain $s_{\mathrm{new}} \in \mathbb{F}_p$ and C_2 to obtain c_{FBE}, then recovers all $\ell + 1$ messages. We bound how much the joint view can depend on b^* by a hybrid over its two components.

FBE component. The $\ell + 1$ messages are embedded in c_{FBE} and the corresponding keys are disclosed. By Lemma 5, the joint distribution $(c_{\mathsf{FBE}}, \{\mathsf{SK}_{\mathsf{FBE},i}\}_{i=0}^{\ell})$ is within $\varepsilon_{\mathrm{FBE}}(\lambda)$ of being invariant under the permutation that swaps which message is designated true; in particular the key dk_{b^*} produced by Enc is within $\varepsilon_{\mathrm{FBE}}$ of being identically distributed to a key produced by $\mathsf{AddDecoy}$. Hence the FBE component changes the adversary's success probability by at most $\varepsilon_{\mathrm{FBE}}$ across values of b^*.

IE component. By Lemma 3, conditioned on the adversary's view (which contains T and all candidate share values but not x_0), the distribution of s_{new} is within $\delta(\lambda)$ of uniform and, up to $2\delta(\lambda)$, independent of which message m_{b^*} was placed at $P(0)$. The adversary cannot link s_{new} to any particular message without x_0: doing so would require identifying the seed, and by the coverage law the number of seeds consistent with the fixed transcript is, up to δ, the same for every candidate message, so no message is distinguished by its preimage count. Hence the IE component changes the adversary's success probability by at most $2\delta(\lambda)$.

Cover-concentration component. Lemma 3 handles the generic case, but it is stated over the randomness of the seed-derived abscissae. It says nothing about the event that the ordinates themselves collapse, which is where a repetition-heavy cover bites. Lemma 4 isolates that event and bounds its contribution by $f_{\max}(\tau)^{k-1} = 2^{-(k-1)H_\infty(\tau)}$, negligible by condition 4 of Definition 6.

Combining. The IE layer (producing s_{new}) and the FBE layer (producing c_{FBE} and keys) use independent randomness, so the deviations add. The adversary's advantage over the $1/(\ell + 1)$ guessing baseline is therefore at most $2\delta(\lambda) + 2^{-(k-1)H_\infty(\tau)} + \varepsilon_{\mathrm{FBE}}(\lambda)$, which is $\mathrm{negl}(\lambda)$ under the stated seed-entropy condition and over admissible covers. □

Lemma 4 (Cover degeneracy, and where it bites). *Work in the random oracle model over a cover τ satisfying conditions 1–3 of Definition 6, and let $f_{\max} = f_{\max}(\tau)$. Let E be the event that all $k - 1$ seed-selected shares carry a single common value $v \in \mathbb{F}_p$. Then $\Pr[E] \le f_{\max}^{k-1}$, and the CD advantage acquires an additive term of at most*

$$\Pr[E] \;\le\; f_{\max}(\tau)^{\,k-1} \;=\; 2^{-(k-1)H_\infty(\tau)},$$

which condition 4 of Definition 6 makes negligible.

Proof. The seed drives an independent selection of each of the $k-1$ share positions via the PRNG modelled as a random oracle, so for a fixed v the chance that one selection lands on a position carrying v is the frequency of v in the cover, at most $f_{\max}$; independence across the $k-1$ selections gives $f_{\max}^{k-1}$, and the events for distinct v are disjoint, so the union over v is still bounded by $\max_v(\cdot) \le f_{\max}^{k-1}$.

It remains to see why E is the bad event. Write $\Phi_m(x_0) = m \cdot L_0(x_{\text{new}}) + A(x_0)$ with $A(x_0) = \sum_{j=1}^{k-1} s_j L_j(x_{\text{new}})$, and recall that inverting the transcript gives the induced candidate $m(x_0) = (s_{\text{new}} - A(x_0)) \cdot L_0(x_{\text{new}})^{-1}$. Off E, the seed feeds *two* independent sources of variation into this expression, the abscissae (through L_0 and the L_j) and the selected ordinates (through A), and the counting argument of Lemma 3 applies unchanged. On E the second source vanishes: with every s_j equal to v and $\sum_{j=0}^{k-1} L_j(x_{\text{new}}) = 1$, the constant term collapses to $A = v\big(1 - L_0(x_{\text{new}})\big)$, so

$$m(x_0) \;=\; v + (s_{\text{new}} - v) \cdot L_0(x_{\text{new}})^{-1}$$

is a deterministic function of the single quantity $L_0(x_{\text{new}})$, whose pushforward over $\mathbb{F}_p$ is not uniform. The preimage counts are therefore unbalanced on E, and bounding the resulting advantage by $\Pr[E]$ gives the claim. □

Remark 9 (What the degenerate event does, and what it does not do). It is tempting to read Lemma 4 as saying that the candidate message $m = v$ acquires an anomalously large anonymity set. That reading is wrong, and it is worth recording because the correct mechanism is less obvious. On E the expression above forces $m = v$ only when $s_{\text{new}} = v$, which by $\Phi_{m^*}(x_0) = v + (m^* - v)L_0(x_{\text{new}})$ happens exactly when the true message already equals v. The failure is not that one designated candidate gains weight; it is that the induced-message map loses one of its two independent randomness sources and stops being balanced across $\mathbb{F}_p$ as a whole. A simulation of the construction over $\mathbb{F}_{1009}$ at $k = 5$ bears this out: the statistical distance of the induced-message distribution from uniform sits at the sampling-noise floor for a lexically diverse cover ($f_{\max} = 0.005$) and rises to 0.13, 0.25 and 0.34 at $f_{\max} = 0.75$, 0.90 and 0.975, while the candidate $m = v$ is *under*-represented throughout. In every resolvable configuration the measured distance stayed below $f_{\max}^{k-1}$, so the bound of the lemma is loose rather than tight, and condition 4 is sufficient rather than necessary.

Remark 10 (Why derive shares from the cover rather than from the seed?). A natural alternative is to derive the $k-1$ shares as $\mathsf{PRG}(x_0)$ outputs, making their distribution uniform by fiat and dispensing with condition 4 altogether. This collapses the construction. If the shares were PRG outputs, s_{new} would be a function of (x_0, m^*) alone, that is, an ordinary symmetric encryption of m^* dressed up as a share, and the cover text would be cryptographically inert decoration. The scheme would no longer be steganographic in any meaningful sense, because the secret would not be a function of the public transcript. Deriving shares from T makes the cover the carrier of the key material: recovery requires

(x_0, T, s_{new}) jointly, the transcript itself is the codebook, and the construction gains a second, independent hardness source, namely the $2^{-\Theta(L)}$ combinatorics of guessing which word subset carries the shares, on top of the $2^{-(\lambda+s)}$ seed entropy [3]. The two barriers degrade separately: partial seed leakage weakens one without touching the other (Section 7.1). The price of using public, non-uniform share values instead of hidden uniform ones is exactly that the CD bound becomes statistical rather than perfect and acquires the concentration term above, which is why Definition 6 has to control it and Theorem 4 has to carry it.

Lemma 5 (FBE permutation symmetry). *Let c_{FBE} embed $\ell+1$ pairwise-distinct messages with keys $\{\mathsf{SK}_{\mathsf{FBE},i}\}_{i=0}^{\ell}$, where the index-$b^*$ message is inserted by* Enc *and the rest by* AddDecoy*. If* Enc *and* AddDecoy *produce keys and container updates from the same distribution (which holds for the FBE construction of [1], where every message contributes one fresh field element and a key specifying which container and key-base indices reconstruct it, with insertion order recorded only in the secret keys and not in c_{FBE}), then the joint distribution $(c_{\mathsf{FBE}}, \{\mathsf{SK}_{\mathsf{FBE},i}\}_{i=0}^{\ell})$ is invariant under any permutation of which index is designated true, so $\varepsilon_{\mathrm{FBE}} = 0$. Any residual asymmetry in a concrete instantiation (for example, a container-growth pattern that correlates with insertion order) is captured by the advantage $\varepsilon_{\mathrm{FBE}}$, which a deployment must keep negligible by inserting all $\ell+1$ messages before release and randomizing container layout.*

Proof. The container $c_{\mathsf{FBE}} \in \mathbb{F}_p^N$ is a vector of field elements; by the construction of [1] each message is represented by an underdetermined linear relation over a shared key-base, and the element appended for the true message is computed by the same rule as for a decoy. Thus the marginal of c_{FBE} is the same regardless of which message was inserted first, and each $\mathsf{SK}_{\mathsf{FBE},i}$ is a tuple of indices into c_{FBE} and the key-base with no field marking it as the Enc-produced key. A permutation of the designation therefore induces the identity on the disclosed joint distribution, giving $\varepsilon_{\mathrm{FBE}} = 0$ in the idealized construction; the deployment caveat is as stated. □

Remark 11 (Why the bound is not exactly zero). The earlier claim $\varepsilon_{\mathrm{CD}} = 0$ held only in an idealization that treated the candidate share values as hidden uniform randomness, which they are not (they are public cover-text word hashes). Once the hiding is correctly attributed to the secret seed, the guarantee becomes statistical and finite: it is exactly zero only in the limit of unbounded seed entropy and a perfectly symmetric FBE container, and otherwise is the explicit negligible quantity $2\delta(\lambda) + \varepsilon_{\mathrm{FBE}}(\lambda)$. This is consistent with IE's own computational (not information-theoretic) posture [3]: the adversary holds all shares and lacks only the seed that says which are used.

6.3 Second-Order Deniability

Corollary 1 (PhantomCrypt is second-order deniable at the network and coercion layers). *When instantiated with an IND-CCA secure PQC-KEM and an IND-CPA secure AEAD scheme, and with seed entropy $\lambda \geq \log_2 p + s$,*

PhantomCrypt is second-order deniable in the sense of Definition 9, with $\varepsilon_{\mathrm{ED}} \leq \varepsilon_{\mathrm{KEM}} + 2\varepsilon_{\mathrm{AEAD}} = \mathit{negl}(\lambda)$ *(Theorem 3) and* $\varepsilon_{\mathrm{CD}} \leq 2\delta(\lambda) + 2^{-(k-1)H_\infty(\tau)} + \varepsilon_{\mathrm{FBE}}(\lambda) = \mathit{negl}(\lambda)$ *in the random oracle model over an admissible cover (Theorem 4). By Lemma 1, a two-stage adversary who first observes the network and then coerces for keys succeeds with probability at most* $\frac{1}{2(\ell+1)} + \mathit{negl}(\lambda)$*. This holds against the network observer and the coercer; it does not cover a forensic adversary who inspects the device (Section 1.1).*

6.4 Post-Quantum Security and Multi-Message Use

PhantomCrypt's security against quantum adversaries rests on different footings for its two halves, and we are careful not to overstate the composed claim. ED reduces to KEM and AEAD security, which are post-quantum when instantiated with ML-KEM-768 (IND-CCA under Module-LWE [33]) and a standard AEAD; this half is post-quantum in the usual sense. The CD half is proved in the (classical) random oracle model. We do *not* claim that this proof transfers to the quantum-accessible random oracle model (QROM) for free: a quantum adversary may query H in superposition, and ROM arguments of the kind used in Lemma 3 and Theorem 4 do not automatically lift to QROM. FBE's underlying information-theoretic symmetry (Lemma 5) holds against unbounded adversaries and is unaffected, but the seed-richness/coverage argument that controls $\delta(\lambda)$ is stated classically. A formal end-to-end QROM proof of composed CD is therefore left as an explicit open problem, not asserted; we expect it to hold but do not prove it here. This asymmetry is why the title speaks of a post-quantum *envelope* rather than of post-quantum deniability wholesale: the quantum-hard assumption sits in the wrapper, while the hardness on the CD side is statistical (seed entropy and cover spread) and therefore involves no assumption a quantum adversary could break, but also no assumption whose ROM analysis lifts to the QROM for free.

Within ML-KEM, the choice of the 768 parameter set is driven by ED rather than by generic strength, and the reason is worth stating because it does not generalise to other components. The ED guarantee is indistinguishability from a reference distribution, and a reference distribution is only useful if it is a protocol that plausibly occurs in the observed environment. ML-KEM-768 is the parameter set deployed by default in hybrid TLS key exchange by major browsers and CDNs, which makes it the natural reference object. ML-KEM-512 (NIST category 1, 768-B ciphertext) would save 320 B but is rare in deployments, so a reference built on it would itself be anomalous. ML-KEM-1024 (category 5, 1568-B ciphertext) would add 480 B for strength that the statistical CD side cannot match anyway, and is likewise uncommon on the wire. ML-KEM-768 (category 3, 1088-B ciphertext) balances the two and, decisively for ED, is what standard traffic actually carries. Swapping parameter sets changes only $\varepsilon_{\mathrm{KEM}}$ and the wire length, and it changes them in the construction and in $\mathcal{D}_{\mathrm{ref}}$ simultaneously, so the reduction of Theorem 3 is unaffected.

Single-message deniability does not automatically extend to multi-message use: naive reuse of the seed x_0 across messages leaks the pairwise message dif-

ferences. We give the concrete attack, a seed-ratcheting mitigation, and the corresponding security statement in Appendix B.

7 A Feasibility Threshold for Coercion Deniability

This section gives the full analysis promised in Section 1, not a sketch: a coverage law (Theorem 5), a concentration bound (Lemma 6), a feasibility direction above the threshold (Theorem 6), and a matching information-theoretic impossibility result below it (Theorem 7), with the critical point made precise in Remark 12.

Theorem 4 bounds the CD advantage by $2\delta(\lambda) + \varepsilon_{\mathrm{FBE}}$ and asserts that $\delta(\lambda)$ vanishes once the seed entropy exceeds $\log_2 p$ by a margin. We now make that dependence exact and show it is a genuine threshold rather than a proof artifact: *undetectability and strong (identical-distribution) coercion deniability are coupled through a single quantity, the seed-entropy margin, and below a sharp line strong deniability is information-theoretically impossible.* Detection-only analyses of IE never see this line because they do not require a fake to be openable; it appears only when both axes are demanded at once.

Setup. Work in the random-oracle model. Fix the field $\mathbb{F}_p$, a seed space $\mathcal{S} = \{0,1\}^\lambda$ with $S := |\mathcal{S}| = 2^\lambda$, a threshold $k \geq 3$, and a public cover text T. For a *fixed observed transcript* (T, s_{new}) define the decryption map

$$\Phi_{T,s_{\mathrm{new}}} : \mathcal{S} \to \mathbb{F}_p, \qquad \Phi_{T,s_{\mathrm{new}}}(x_0) = \mathsf{Dec}(x_0, T, s_{\mathrm{new}}),$$

sending a seed to the message it reconstructs from that transcript. (By the argument of Section 5, Dec is total on candidate seeds and every seed yields a forward-consistent claim, so no seed can be rejected by the coercer on consistency grounds.) For $m \in \mathbb{F}_p$ the *anonymity set* is $\Phi^{-1}(m) = \{x_0 : \Phi(x_0) = m\}$. Let the user's fake messages be drawn from a *plausible source* $Z \sim \mathcal{P}$ over a support $\mathrm{supp}(\mathcal{P}) \subseteq \mathbb{F}_p$ with Shannon entropy $H(Z) = H(\mathcal{P})$; for the strong identical-distribution guarantee the real message is drawn from the same Z. Define the *seed-entropy margin*

$$s \;:=\; \lambda - H(\mathcal{P}), \qquad \text{equivalently} \qquad \rho \;:=\; S/2^{H(\mathcal{P})} \;=\; 2^s.$$

Throughout this section we work in the strong identical-distribution regime with a *near-uniform* plausible source, $H(\mathcal{P}) \approx H_0(\mathcal{P}) \approx \log_2 p$ — the natural model when the plausible payloads are themselves high-entropy (key material, document hashes, coordinate blobs), as in the whistleblower scenario of Section 8. In this regime the seed-entropy margin coincides with the field-size margin,

$$s \;=\; \lambda - H(\mathcal{P}) \;=\; \lambda - \log_2 p, \qquad \text{so that} \qquad \rho \;=\; 2^s \;=\; S/p,$$

and every coverage statement below is governed by the single quantity s; this recovers the field-size form of Theorem 4. The connection to FBE's plausibility distance [1] is that $H(\mathcal{P})$ plays the role of the conditional plaintext entropy there.

For a genuinely low-entropy decoy source the coupling splits into two distinct margins — a field-size margin $\lambda - \log_2 p$ governing coverage and an entropy margin $\lambda - H(\mathcal{P})$ governing the displayed-message reweighting — which we do not treat here and flag as future work (Section 9).

Theorem 5 (Coverage law). *Model $\Phi_{T,s_{\mathrm{new}}}$ as a random function (the random-oracle idealization of the seed-driven index selection and abscissae). Then for any fixed $m \in \mathbb{F}_p$ the anonymity-set size $|\Phi^{-1}(m)|$ is distributed* Binomial$(S, 1/p)$*, hence has mean S/p and is approximated by* Poisson(S/p) *for large S, up to a negl(λ) correction from the negl(λ) fraction of seeds discarded for abscissa collisions. Consequently the expected fraction of the plausible support reachable by some seed is*

$$\mathbb{E}\left[\frac{|\Phi(\mathcal{S}) \cap \mathrm{supp}(\mathcal{P})|}{|\mathrm{supp}(\mathcal{P})|}\right] \;=\; 1 - e^{-S/p}.$$

In words: S/p is simply how many seeds the construction supplies per field element, and a given target message is openable precisely when at least one seed lands on it, which happens with probability $1 - e^{-S/p}$. So $S/p \gg 1$ means almost every candidate message can be reached by some seed (deniability is feasible), while $S/p \ll 1$ means almost none can (it is not). The coverage is thus controlled by the seed-to-field ratio $S/p = 2^{\lambda - \log_2 p}$; this is the field-size form, and it equals the seed-entropy form 2^s exactly when $\mathcal{P}$ is uniform over $\mathbb{F}_p$ (so that $H(\mathcal{P}) = \log_2 p$).

Proof. For a random function $\Phi : \mathcal{S} \to \mathbb{F}_p$ and fixed m, each seed maps to m independently with probability $1/p$, so $|\Phi^{-1}(m)| \sim$ Binomial$(S, 1/p)$ with mean S/p; the Poisson limit is standard. A seed is discarded only if its k hash-derived abscissae fail to be distinct and nonzero, which by the birthday bound over $\mathbb{F}_p$ happens for a fraction $O(k^2/p) =$ negl(λ) of seeds, contributing the stated correction. The probability that a fixed m is *unreachable* is $\Pr[\text{Binomial}(S, 1/p) = 0] = (1 - 1/p)^S \to e^{-S/p}$, so the expected reachable fraction is $1 - e^{-S/p}$. □

Lemma 6 (Anonymity-set concentration). *Under the model of Theorem 5, for every m with $\Phi^{-1}(m) \neq \emptyset$ and any $\beta \in (0, 1)$,*

$$\Pr\big[\big||\Phi^{-1}(m)| - S/p\big| > \beta\, S/p\big] \;\le\; 2\exp\Big(-\tfrac{\beta^2}{3} \cdot \tfrac{S}{p}\Big).$$

In particular, when $S/p = 2^s$ all nonempty anonymity sets have size $(1 \pm o(1))S/p$ except with probability $2^{-\Omega(2^s)}$.

Proof. A multiplicative Chernoff bound on the Binomial$(S, 1/p)$ variable with mean $\mu = S/p$. □

Theorem 6 (Feasibility above the threshold). *Suppose the seed-entropy margin satisfies $s \ge \omega(\log \lambda)$, i.e. $\rho = 2^s = S/p$ is super-polynomial (under the near-uniform standing assumption this is just $\lambda \ge \log_2 p + \omega(\log \lambda)$). Then, except with probability negl(λ) over the random oracle:*

1. (Openability.) *every plausible message $m \in \mathrm{supp}(\mathcal{P})$ has a nonempty anonymity set, so for any freely chosen fake m' the user can surrender a seed opening the observed transcript to m'; and*
2. (Indistinguishability.) *the coercer's view $(T, s_{\mathrm{new}}, x_0^{\mathrm{disc}})$ under an honest run with real message $m^* \sim Z$ is statistically close to its view under a coerced run in which the user surrenders a uniformly chosen seed from $\Phi^{-1}(m')$ for a fake $m' \sim Z$; the statistical distance is $2^{-\Omega(s)}$.*

Hence PhantomCrypt is single-coercion plausibly deniable in the strong (identical-distribution) sense, and the bound $\delta(\lambda)$ of Lemma 3 satisfies $\delta(\lambda) \le 2^{-\Omega(s)}$.

Proof. Openability. By Theorem 5 a fixed m is unreachable with probability $e^{-S/p}$. The plausible support has size $|\mathrm{supp}(\mathcal{P})| \le 2^{H_0(\mathcal{P})}$, with H_0 the max-entropy (log-support-size). A union bound over the support gives

$$\Pr[\exists\, m \in \mathrm{supp}(\mathcal{P}) : \Phi^{-1}(m) = \emptyset] \;\le\; 2^{H_0(\mathcal{P})}\, e^{-S/p}.$$

Since $S/p = 2^s$, the hypothesis $s \ge \omega(\log \lambda)$ gives $S/p = \lambda^{\omega(1)}$ (super-polynomial), so the bound is at most $2^{H_0(\mathcal{P})} \exp(-\lambda^{\omega(1)}) = \mathrm{negl}(\lambda)$: the super-polynomial S/p in the exponent dominates the $H_0(\mathcal{P}) \le \log_2 p$ bits of the support. Thus all plausible targets are simultaneously reachable w.h.p.

Indistinguishability. Condition on the (overwhelmingly likely) event that all anonymity sets are nonempty and, by Lemma 6, of size $(1 \pm \beta)S/p$ with $\beta = 2^{-\Omega(s)}$. In the honest run, $m^* \sim Z$ and x_0^{disc} is the real seed, which is uniform on $\mathcal{S}$ and induces $m^* = \Phi(x_0^{\mathrm{disc}})$; equivalently x_0^{disc} is uniform on $\Phi^{-1}(m^*)$ with m^* marginally $\sim Z$. In the coerced run, $m' \sim Z$ and x_0^{disc} is uniform on $\Phi^{-1}(m')$. The two processes differ only in how the displayed message is sampled relative to the transcript, but in both the surrendered seed is uniform on the anonymity set of a Z-distributed message. The only discrepancy is the reweighting induced by unequal anonymity-set sizes: a message m is displayed in the honest run with probability proportional to $|\Phi^{-1}(m)| \cdot \Pr_Z[m]$ rather than $\Pr_Z[m]$. Since all sizes lie within $(1 \pm \beta)S/p$, this reweighting perturbs the displayed-message distribution by total variation at most $\beta = 2^{-\Omega(s)}$, and conditioned on the displayed message the surrendered seed is identically (uniform on the anonymity set) distributed in both runs. Hence the coercer's joint view has statistical distance $\le 2^{-\Omega(s)}$ between the two runs. Identifying this distance with $\delta(\lambda)$ gives the final claim. □

Theorem 7 (Impossibility below the threshold; information-theoretic). *Let the coercer be computationally* unbounded. *If the margin satisfies $s \le -\omega(\log \lambda)$, i.e. $\rho = S/2^{H(\mathcal{P})} \le 1/\mathrm{poly}(\lambda)$, then a $1 - o(1)$ fraction of plausible messages have empty anonymity sets:* $\Pr_{m \sim Z}[\Phi^{-1}(m) = \emptyset] \to 1$. *For any such target m' there is* no *seed whatsoever that opens the fixed observed transcript to m', so the user cannot exhibit m' as a plausible content regardless of effort or computational power. Strong (identical-distribution) plausible deniability is therefore unachievable in this regime, against even an unbounded adversary.*

Proof. By Theorem 5, a message m is unreachable with probability $e^{-S/p}$. For $m \sim Z$, $\mathbb{E}_m \Pr[\Phi^{-1}(m) = \emptyset] = e^{-S/p}$; under the near-uniform standing assumption $S/p = 2^s = \rho$, so when $\rho \leq 1/\mathrm{poly}$ the reachable fraction $1 - e^{-S/p} \leq S/p = \rho \leq 1/\mathrm{poly}(\lambda) \to 0$, and the unreachable fraction tends to 1. Reachability is a property of the fixed transcript and the map Φ alone, independent of the adversary's power, so no unbounded strategy by the user can manufacture a seed for an unreachable target. Strong deniability requires opening the user's freely chosen fake; with probability $1 - o(1)$ that fake is unreachable, so the requirement fails. □

Remark 12 (Sharpness and the critical point). The controlling quantity is $\exp(-2^s)$, which moves from ≈ 1 to ≈ 0 as s crosses zero over a window of width $O(\log \lambda)$, so Theorems 6 and 7 meet at a sharp threshold. At the critical point $s = 0$ ($\rho = 1$, the original IE parameterization $\lambda = \log_2 p$ with flat $\mathcal{P}$), the reachable fraction is exactly $1 - e^{-1} \approx 0.632$: about 37% of freely chosen fakes are unopenable, so the as-published parameters do *not* support strong plan-ahead deniability. The fix is to provision the seed with a margin above the field size, $\lambda \geq \log_2 p + s$ for a statistical parameter s (e.g. $s = 64$, giving $S/p = 2^{64}$ and a reachable fraction $1 - 2^{-\Omega(2^{64})}$), at no cost to undetectability, which is the design rule we adopt.

Remark 13 (Computational bridge to the concrete and post-quantum settings). Theorems 5–7 are stated in the random-oracle idealization, where Φ is a random function and the guarantees are statistical. With a concrete hash and a post-quantum carrier (Section 6.4), Φ is only *pseudorandom*: the coverage law and the feasibility direction then hold against computationally bounded distinguishers, with the degradation equal to the hash's PRF/RO-distinguishing advantage (in the quantum setting, its QROM-distinguishing advantage). The impossibility direction (Theorem 7) is unaffected, since unreachability is a combinatorial property of the realized map and holds against unbounded adversaries irrespective of how Φ is instantiated. This is the precise sense in which undetectability is computational while the deniability *obstruction* is information-theoretic.

7.1 Seed distribution, rotation, and partial leakage

The deniable secret is the seed $\mathsf{SK}_{\mathrm{ie}} = (x_0, k)$, and every guarantee above is conditioned on its secrecy, so its lifecycle is a first-class part of the design rather than an appendix footnote.

Distribution. x_0 is pre-shared out of band, never written to the device, and memorizable as a 24-word mnemonic (≈ 256 bits); it is absent from every artefact disclosed under coercion (Section 1.1). This is exactly what makes the "no withheld key" property hold: the coercer receives a complete, self-consistent key set, and x_0 leaves no gap in it to point at.

Rotation. Reusing one x_0 across messages is unsafe it leaks pairwise message differences (Lemma 7). The mitigation is to ratchet, $x_0^{(i+1)} \leftarrow H(x_0^{(i)} \,\|\, C^{(i)})$,

which in the random-oracle model makes per-message seeds computationally independent and restores per-message CD and ED (Theorem 8). We state the rule here because it is a deployment requirement; the attack and proof are in Appendix B.

Partial leakage degrades the margin, not the scheme. A more realistic failure than full disclosure is *partial* leakage, a side channel, a low-entropy mnemonic, or coercion that extracts some seed bits. This maps cleanly onto the threshold of this section. If the adversary learns b bits of x_0, the residual seed space has size $2^{\lambda-b}$, so the operative ratio becomes $S'/p = 2^{(\lambda-b)-\log_2 p} = 2^{s-b}$, and every statement above holds verbatim with the margin s replaced by $s-b$: the coverage law reads $1-\exp(-2^{s-b})$ and the indistinguishability distance reads $2^{-\Omega(s-b)}$. Deniability, therefore, slides *along the same curve* rather than collapsing: it survives as long as $s-b$ stays above the statistical parameter, which is precisely why the design rule provisions a margin (e.g. $s=64$) instead of sitting at the $s=0$ critical point. Full leakage ($b=\lambda$) recovers the known boundary the adversary then holds x_0 and reconstructs m^* which is the Stage-2 case where CD, not ED, is the surviving guarantee.

8 Implementation and Evaluation

We implemented PhantomCrypt in Python; source at `https://anonymous.4open.science/r/PhantomCrypt-D532`. Instantiation: $p = 2^{256} - 2^{32} - 977$, SHA3-256 for H, ML-KEM-768 [33], and AES-256-GCM. Benchmarks on an Intel Core i5-1135G7 at 2.42 GHz (Python 3.10), averaged over 100 runs.

Baseline comparison. The natural baseline is the reference protocol itself: vanilla hybrid PKE (ML-KEM-768 + AES-256-GCM) wrapping a 32-byte payload, which on our hardware costs 0.21 ms to encapsulate-and-encrypt, 0.11 ms to decapsulate, and 1148 B on the wire. PhantomCrypt at $k=5$, $\ell=3$ produces 1464 B on the wire, an overhead of 316 B (+27.5%) over the baseline, dominated by the 1088-B KEM ciphertext that the baseline also carries; the marginal cost of deniability is the FBE container and the second AEAD frame. Encryption of a 32-byte message takes ≈7.4 ms, true decryption ≈7.0 ms, decoy decryption ≈0.7 ms; the gap between PhantomCrypt's 7.4 ms and the baseline's 0.21 ms is FBE container construction in the unoptimized Python proof of concept, not the post-quantum primitives. Overhead is constant in message size.

Scaling. Two parameters drive cost: cover-text length L and decoy count ℓ. The dependence on L is where the threshold choice (Section 3.2) becomes concrete, and we report both regimes rather than only the cheaper one. For *constant* k (closed-world CD only), IE share derivation interpolates through k points regardless of L and is effectively flat in cover length: we measure 0.11 ms at $L=100$ and 0.11 ms at $L=5000$ words with $k=5$. This figure does *not* buy the open-world (t,q) guarantee, which requires $k=\Theta(L)$. In that security-relevant regime the

cost is no longer flat: interpolation through $\Theta(L)$ points and the corresponding share derivation grow with L. Table 3 reports *measured* share-derivation-plus-interpolation time at $k = \lceil L/8 \rceil$ (a representative growing schedule) over the 256-bit field $p = 2^{256} - 2^{32} - 977$; the cost scales super-linearly in L as expected for dense interpolation, reaching ≈100 ms for a 2000-word cover in the unoptimized Python prototype. We report this regime explicitly because benchmarking only the constant-k case would advertise a flat cost for a parameterization that does not support the open-world deniability the paper otherwise claims. Wire overhead grows linearly in ℓ at one field element (32 B) per decoy in the two-payload design (Table 4). This linear growth is itself a mild ED signal, since the ciphertext length then encodes ℓ. Two variants remove it (Table 4, last row): keeping the FBE container off the wire (steg/QR delivery, Section 5) returns the wire to the 1148-B baseline regardless of ℓ, while a single-payload fixed-bucket layout that merges the two AEAD frames and pads to $\ell_{\max} = 10$ holds the wire at a constant 1660 B (overhead 512 B, +224 B of padding at $\ell = 3$), trading a fixed length-independent cost for the elimination of the decoy-count leak.

Table 3: Measured share-derivation-plus-interpolation cost in the security-relevant growing-threshold regime $k = \lceil L/8 \rceil$ (open-world (t, q)-deniability), over $p = 2^{256} - 2^{32} - 977$ with SHA3-256 as H. Contrast with the constant-k figure of 0.11 ms, which supports closed-world CD only. Unoptimized Python prototype, best of three runs.

L (words)	$k = \lceil L/8 \rceil$	derive + interpolate (ms)
100	13	1.89
500	63	12.66
1000	125	33.61
2000	250	99.95

Statistical sanity check against a byte-distribution detector. Our ED guarantee is by reduction (Theorem 3), but a reviewer reasonably asks whether the constructed object exhibits an *obvious* statistical anomaly that a cheap detector would flag, independent of the reduction. We therefore ran the simplest such test: a byte-frequency uniformity check on the on-wire object $(C_{\text{kem}}, C_1, C_2)$, comparing a real PhantomCrypt ciphertext against a draw of $\mathcal{D}_{\text{ref}}$ of the same lengths. Table 5 reports the wire length and a Pearson byte chi-square statistic (256 bins, ≈255 degrees of freedom under uniformity) for both. The two objects have *identical* wire length and chi-square values consistent with uniform bytes for both branches, so a length or byte-histogram detector has no signal to key on. This is consistent with, not a substitute for, the reduction: it rules out gross anomalies but says nothing about traffic shape or timing, which remain out of scope (Remark 2). We deliberately do not claim resistance to a trained distri-

Table 4: Wire overhead versus decoy count ℓ ($k = 5$); KEM ciphertext 1088 B, each $\mathbb{F}_p$ element 32 B. Baseline vanilla hybrid PKE is 1148 B. The last row is the single-payload fixed-bucket variant: the two AEAD frames are merged into one (saving a 12-B nonce and a 16-B tag) and the container is padded to a fixed $\ell_{\max} = 10$ bucket, so the wire length is *constant in* ℓ and no longer encodes the decoy count. Its value is derived from the two-payload measurements above by this accounting, not separately benchmarked.

ℓ	total wire (B)	overhead over baseline (B)
1	1400	252
3	1464	316
5	1528	380
10	1688	540
fixed bucket ($\ell \leq 10$)	1660	512

butional classifier or live DPI; those require a transport-layer study we leave to future work.

Table 5: Byte-distribution sanity check on the on-wire object. Real PhantomCrypt ($k = 5$, $\ell = 3$) versus a length-matched draw of $\mathcal{D}_{\text{ref}}$. Chi-square is the Pearson statistic over 256 byte values ($\approx$255 d.o.f. under uniformity); the near-equal sub-255 values for the two branches indicate no byte-frequency signal. Cover string excluded (identical in both branches).

object	wire length (B)	byte χ^2
PhantomCrypt $(C_{\text{kem}}, C_1, C_2)$	1472	237.9
$\mathcal{D}_{\text{ref}}$ draw (same lengths)	1472	238.6

Cover admissibility on real text. A natural worry is that the admissibility conditions of Definition 6 are brittle for natural language and might force unnatural authoring that itself harms plausibility. We measured all four conditions directly. Taking a corpus of nine real Wikipedia articles (the cover source used in the original IE evaluation), totalling $\approx$103,000 word tokens and ranging from 5,700 to 15,500 words each, we sliced them into consecutive non-overlapping windows of length $L \in \{50, 100, 200, 500, 1000, 2000\}$ and tested the exact predicate (SHA3-256 into $\mathbb{F}_p$, $p = 2^{256} - 2^{32} - 977$) on every window *without any modification to the text*, at both the constant-threshold regime ($k = 5$) and the growing-threshold regime ($k = \lceil L/8 \rceil$, up to $k = 250$).

Conditions 1–3. All windows passed at every length and in both regimes, with no failures on length, distinctness, or share separation, including all forty-seven 2000-word windows matching the whistleblower scenario. The reason is

structural: natural English of even moderate length carries far more than k distinct words, and SHA3 into a 256-bit field makes a share collision a birthday event of probability $O(L^2/p)$, i.e. negligible.

Condition 4. The frequency condition is the interesting one, because natural language is Zipfian and therefore *does* concentrate. Across all forty-seven windows at $L = 2000$ the worst-case dominant token (typically *the*) occupied $f_{\max} \leq 9.1\%$ of positions, giving $H_\infty(\tau) \geq 3.5$ bits per selection with an average of 4.1 bits; the worst case over all 3,960 windows at every length was $f_{\max} = 24\%$ at $L = 50$, i.e. $H_\infty(\tau) \geq 2.06$ bits. Substituting into the concentration exponent $(k-1)H_\infty(\tau)$ of Lemma 4 separates the two regimes sharply.

- *Growing threshold* ($k = \lceil L/8 \rceil$): the exponent is ≥ 31 bits already at $L = 100$ ($k = 13$) and ≥ 863 bits at $L = 2000$ ($k = 250$). Condition 4 is satisfied by an enormous margin and imposes no authoring constraint whatsoever.
- *Constant threshold* ($k = 5$): the exponent is only 8 to 14 bits, so the term $f_{\max}^{k-1}$ sits between 2^{-8} and 2^{-14}. A constant does not become negligible in λ, so at constant k the bound of Theorem 4 does not close asymptotically.

Two cautions on how to read the constant-k number, both of which cut against overstating it. First, it is a statement about the *bound*, not a measured bias: Remark 9 reports that the actual statistical distance stays well below $f_{\max}^{k-1}$ wherever it is resolvable, so condition 4 is sufficient rather than necessary and the true advantage at $f_{\max} \approx 0.09$ is plausibly far smaller. Second, and more importantly, the concentration term is not the binding constraint at constant parameters anyway: by Theorem 4 the bound is negligible only when $s = \omega(\log \lambda)$, so a deployment that fixes s has already forfeited the asymptotic guarantee before the cover is considered at all. What the measurement does establish is that natural text supplies ample margin in the growing-threshold regime, and that the frequency condition is the one admissibility requirement natural language does not satisfy for free. Small constant thresholds also forfeit the open-world (t, q) guarantee (the constant-threshold caveat of [3]). A deployment that wants the full asymptotic guarantee should run in the growing-threshold regime, whose cost is reported in Table 3. Condition 4 is efficiently checkable from τ alone by counting token frequencies, and a cover that fails it should be lengthened or replaced rather than edited, since editing to raise H_∞ is exactly the kind of unnatural authoring that harms plausibility on the ED side.

Constant-time considerations. Deniability is a confidentiality property and is undone by timing leakage. The security-sensitive operations are $\mathbb{F}_p$ arithmetic, Lagrange interpolation, and PRNG-driven word selection; all admit constant-time implementation, and the share count k is fixed and public. The Python proof of concept is *not* constant-time (it uses arbitrary-precision integers with data-dependent timing), and a deployment must use constant-time field arithmetic and avoid input-dependent branches in selection; we flag this as an implementation requirement rather than a property of the scheme.

Kerckhoffs-consistency. $\mathsf{SK}_{\mathrm{ie}} = (x_0, k)$ is pre-shared, memorizable (24-word mnemonic), and destroyable; CD survives even if sk_{kem} and all FBE keys leak, with the negligible advantage of Theorem 4. If PhantomCrypt software itself is found on a seized device (Stage 2 of Section 1.1), 2OD no longer holds, but CD remains by Theorem 4.

Application scenarios. Activists embed GPS coordinates within an academic paper (cover text), with five decoy messages for plausible cover. A whistleblower embeds a 32-byte document hash within a 2000-word article; under coercion, a decoy FBE key produces "Meeting confirmed."

9 Conclusion

We argued that second-order deniability, the deniability of using a deniability mechanism, is best treated not as one primitive but as the conjunction of two formally separate and logically independent properties: content deniability and existence deniability. Both rest on threshold secret sharing over $\mathbb{F}_p$, and PhantomCrypt composes FBE (CD) with IE (ED) in a hybrid PQC-KEM envelope so that standard encryption serves as its own cover. We were deliberate about the limits. ED is indistinguishability of the ciphertext *object* from a fixed reference distribution, conditional on the cover string being unremarkable and excluding traffic-analysis resistance; it reduces to KEM and AEAD security. the CD advantage is negligible in the random oracle model, with the residual advantage controlled by the seed entropy relative to $\log_2 p$ and by the FBE container's permutation symmetry, rather than exactly zero. The open-world (t, q) guarantee requires a threshold growing with the cover length, a cost we benchmark honestly rather than report only the flat constant-k case. And 2OD in the full sense fails once the mechanism's software is found on a seized device, at which point CD still protects the message.

The most consequential open problems are the ones the present design does not solve. (i) *Software-forensic deniability*: making the presence of the mechanism on a device unremarkable, e.g., by embedding it in a widely deployed messaging or TLS stack, which would close the Stage-2 gap and is the main thing separating PhantomCrypt from anamorphic constructions. (ii) *Quantitative cover plausibility*: ED ultimately rests on the cover text being unremarkable to the observer, and our ED game holds that plausibility fixed across both branches rather than certifying it; a metric for it deserves a model rather than an assumption, and it is the substance of existence deniability that our proofs do not reach. (iii) A formal QROM proof of composed CD, which we conjecture but do not establish, since the seed-richness argument is stated classically and ROM does not lift to QROM for free. Further directions: multi-message CPA/CCA definitions capturing active cross-session adversaries, a feasibility-threshold characterization relating seed entropy, field size, and the number of plausible openings, and usability studies.

Hybridizing with anamorphic encryption. The orthogonality of Table 2 points to a concrete composition that addresses open problem (i) directly. PhantomCrypt's one weakness is Stage-2 software forensics; anamorphic encryption's are capacity and opening symmetry. Carrying PhantomCrypt's transmitted share s_{new} (and, in the on-wire layout, the FBE container) inside the anamorphic channel of a widely deployed PKE scheme would let the deniable payload ride within standard, already-installed encryption software, removing the discoverable non-standard tool that collapses 2OD at Stage 2, while retaining PhantomCrypt's $\ell+1$ structurally symmetric openings and super-polynomial residual uncertainty under full-key disclosure. The anamorphic double key would supply part of sk_{den}. Two questions decide whether this is more than a slogan: whether the anamorphic capacity bounds of [14] admit a payload as large as an FBE container, and whether the composition preserves the ED and CD games simultaneously rather than trading one for the other. We leave both open.

Ethical considerations. Second-order deniability raises dual-use concerns. PhantomCrypt is designed for journalists, activists, and whistleblowers in authoritarian settings, but could be exploited by malicious actors. We follow Albrecht et al. [4] in arguing that rigorous security analysis for vulnerable populations is both a technical and ethical imperative.

References

1. Ahmad, S., Rass, S., Schartner, P.: False-Bottom Encryption: Deniable Encryption From Secret Sharing. IEEE Access **11**, 62549–62564 (2023). https://doi.org/10.1109/ACCESS.2023.3288285, https://ieeexplore.ieee.org/document/10158701/
2. Ahmad, S., Rass, S., Seyedi, Z.: Plausible deniability in fully homomorphic computation (2026), https://arxiv.org/abs/2605.01985
3. Ahmad, S., Rass, S., Zahra, S.: Invisible encryption. In: Data Privacy Management (DPM 2025), 20th International Workshop co-located with ESORICS 2025. pp. 96–111. Toulouse, France (Sep 25 2025), https://link.springer.com/book/9783032160881, pre-print in DPM 2025 Pre-proceedings
4. Albrecht, M.R., Colombo, S., Dowling, B., Jensen, R.B.: At-compromise security: The case for alert blindness. Cryptology ePrint Archive, Paper 2026/252 (2026), https://eprint.iacr.org/2026/252
5. An, Z., Tian, H., Chen, C., Zhang, F.: Deniable Cryptosystems: Simpler Constructions and Achieving Leakage Resilience. In: Computer Security – ESORICS 2023. pp. 24–44. Springer Nature Switzerland, Cham (2024). https://doi.org/10.1007/978-3-031-50594-2_2
6. An, Z., Zhang, F.: Deny whatever you want: Dual-deniable public-key encryption (2025), https://eprint.iacr.org/2025/182
7. Angel, Y.: obfs4 (the obfourscator). Tor Project (2014), https://gitweb.torproject.org/pluggable-transports/obfs4.git
8. Anzuoni, E., Gagliardoni, T.: Shufflecake: Plausible deniability for multiple hidden filesystems on linux. In: Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security. p. 3033–3047. CCS '23, New York, NY, USA (2023), https://doi.org/10.1145/3576915.3623126

9. Banfi, F., Gegier, K., Hirt, M., Maurer, U., Rito, G.: Anamorphic Encryption, Revisited. In: Advances in Cryptology – EUROCRYPT 2024. pp. 3–32. Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-58723-8_1
10. Canetti, R., Damgård, I., Kolby, S., Ravi, D., Yakoubov, S.: Deniable secret sharing. In: Theory of Cryptography: 23rd International Conference, TCC 2025, Aarhus, Denmark, December 1–5, 2025, Proceedings, Part II. p. 399–427. Springer-Verlag, Berlin, Heidelberg (2025), https://doi.org/10.1007/978-3-032-12293-3_13
11. Canetti, R., Park, S., Poburinnaya, O.: Fully Deniable Interactive Encryption. In: Advances in Cryptology – CRYPTO 2020. pp. 807–835. Springer International Publishing. https://doi.org/10.1007/978-3-030-56784-2_27
12. Canetti, R., Dwork, C., Naor, M., Ostrovsky, R.: Deniable Encryption. In: Advances in Cryptology — CRYPTO '97. pp. 90–104. Springer, Berlin, Heidelberg. https://doi.org/10.1007/BFb0052229
13. Carnemolla, D., Catalano, D., Giunta, E., Migliaro, F.: Anamorphic Resistant Encryption: the Good, the Bad and the Ugly. In: Advances in Cryptology – CRYPTO 2025. pp. 472–503. Springer Nature Switzerland. https://doi.org/10.1007/978-3-032-01881-6_15
14. Catalano, D., Giunta, E., Migliaro, F.: Limits of Black-Box Anamorphic Encryption. In: Advances in Cryptology – CRYPTO 2024. pp. 352–383. Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-68379-4_11
15. Catalano, D., Giunta, E., Migliaro, F.: Anamorphic encryption: New constructions and homomorphic realizations. In: Advances in Cryptology – EUROCRYPT 2024: 43rd Annual International Conference on the Theory and Applications of Cryptographic Techniques, Zurich, Switzerland, May 26–30, 2024, Proceedings, Part II. p. 33–62. Springer-Verlag, Berlin, Heidelberg (2024), https://doi.org/10.1007/978-3-031-58723-8_2
16. Chan, C.K., Cheng, L.M.: Hiding data in images by simple LSB substitution. Pattern Recognition **37**(3), 469–474 (Mar 2004), https://www.sciencedirect.com/science/article/pii/S003132030300284X
17. Chang, C.Y., Clark, S.: Practical linguistic steganography using contextual synonym substitution and a novel vertex coding method. Comput. Linguist. **40**(2), 403–448 (Jun 2014), https://doi.org/10.1162/COLI_a_00176
18. Chapman, M., Davida, G.I.: Plausible deniability using automated linguistic stegonagraphy. In: Proceedings of the International Conference on Infrastructure Security. p. 276–287. InfraSec '02, Springer-Verlag (2002)
19. Choi, W., Collins, D., Liu, X., Zikas, V.: A unified treatment of anamorphic encryption. Cryptology ePrint Archive, Paper 2025/309 (2025), https://eprint.iacr.org/2025/309
20. Collins, D., Colombo, S., Huguenin-Dumittan, L.: Real-world deniability in messaging (2023), https://eprint.iacr.org/2023/403
21. Deo, A., Libert, B.: Fully asymmetric anamorphic homomorphic encryption from LWE. Cryptology ePrint Archive, Paper 2025/328 (2025), https://eprint.iacr.org/2025/328
22. Ding, J., Chen, K., Wang, Y., Zhao, N., Zhang, W., Yu, N.: Discop: Provably secure steganography in practice based on "distribution copies". In: 2023 IEEE Symposium on Security and Privacy (SP). pp. 2238–2255 (2023). https://doi.org/10.1109/SP46215.2023.10179287
23. Dodis, Y., Goldin, E.: Anamorphic-Resistant Encryption; Or Why the Encryption Debate is Still Alive. In: Advances in Cryptology – CRYPTO 2025. pp. 440–471. Springer Nature Switzerland. https://doi.org/10.1007/978-3-032-01881-6_14

24. Fenske, E., Johnson, A.: Security notions for fully encrypted protocols. In: FOCI (2023)
25. Fenske, E., Johnson, A.: Bytes to schlep? use a FEP: Hiding protocol metadata with fully encrypted protocols. In: ACM CCS (2024), arXiv:2405.13310
26. Fiedler, R., Janson, C.: A deniability analysis of Signal's initial handshake PQXDH. In: Proceedings on Privacy Enhancing Technologies (PoPETs) (2024). https://doi.org/10.56553/popets-2024-0148, cryptology ePrint Archive, Paper 2024/741
27. Günther, F., Stebila, D., Veitch, S.: Obfuscated key exchange. In: ACM CCS (2024), cryptology ePrint Archive, Paper 2024/1086
28. Irwin, A., Hunt, R.: Forensic methods for detection of deniable encryption in mobile networks (2009). https://doi.org/10.1109/PACRIM.2009.5291379
29. Johnson, N.F., Duric, Z., Jajodia, S.: Information Hiding: Steganography and Watermarking—Attacks and Countermeasures. Kluwer Academic (2001)
30. Juels, A., Ristenpart, T.: Honey encryption: Security beyond the brute-force bound. In: EUROCRYPT. LNCS, vol. 8441, pp. 293–310. Springer (2014)
31. Kaptchuk, G., Jois, T.M., Green, M., Rubin, A.D.: Meteor: Cryptographically secure steganography for realistic distributions. In: Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security. p. 1529–1548. CCS '21, Association for Computing Machinery, New York, NY, USA (2021). https://doi.org/10.1145/3460120.3484550, https://doi.org/10.1145/3460120.3484550
32. Leiken, J., Park, S.: On the credibility of deniable communication in court (2026). https://doi.org/10.1145/3788646.3789520, https://doi.org/10.1145/3788646.3789520
33. National Institute of Standards and Technology: FIPS 203: Module-lattice-based key-encapsulation mechanism standard (ML-KEM). Tech. rep., NIST (2024)
34. Persiano, G., Phan, D.H., Yung, M.: Anamorphic encryption: Private communication against a dictator. In: EUROCRYPT. LNCS, vol. 13276, pp. 34–63. Springer (2022)
35. Persiano, G., Phan, D.H., Yung, M.: Public-key anamorphism in (CCA-secure) public-key encryption and beyond. In: CRYPTO 2024. LNCS (2024), cryptology ePrint Archive, Paper 2024/1327
36. Unger, N., Dechand, S., Bonneau, J., Fahl, S., Perl, H., Goldberg, I., Smith, M.: SoK: Secure messaging. In: 2015 IEEE Symposium on Security and Privacy (SP). pp. 232–249. IEEE (2015). https://doi.org/10.1109/SP.2015.22
37. Unger, N., Goldberg, I.: Deniable key exchanges for secure messaging. In: Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security (CCS). pp. 1211–1223. ACM (2015). https://doi.org/10.1145/2810103.2813616
38. VeraCrypt. https://www.veracrypt.fr/ (2025), accessed 2025
39. Wayner, P.: Disappearing Cryptography: Information Hiding: Steganography and Watermarking. Morgan Kaufmann, 2nd edn. (2002)
40. Winter, P., Pulls, T., Fuss, J.: ScrambleSuit: A polymorphic network protocol to circumvent censorship. In: WPES (2013)
41. de Witt, C.S., Sokota, S., Kolter, J.Z., Foerster, J., Strohmeier, M.: Perfectly secure steganography using minimum entropy coupling. In: ICLR (2023)
42. Wu, M., Sippe, J., Sivakumar, D., Burg, J., Anderson, P., Wang, X., Bock, K., Houmansadr, A., Levin, D., Wustrow, E.: How the Great Firewall of China detects and blocks fully encrypted traffic. In: USENIX Security (2023)

A Correctness of PhantomCrypt

Proposition 2 (Correctness). *PhantomCrypt, under the instantiation* $sk_{\text{den}} = \mathsf{SK}_{\text{ie}} = (x_0, k)$ *and* $dk_i = (sk_{\text{kem}}, \mathsf{SK}_{\mathsf{FBE},i})$*, satisfies Definition 2.*

Proof. (i) Dec *on* $k = sk_{\text{den}}$. Given SK_{ie} and T (the recipient holds sk_{kem} from any dk_i): decapsulate C_{kem} to K_{session}; decrypt C_1 to s_{new}; re-derive $\{(x_j, s_j)\}_{j=1}^{k-1}$ from T via $\text{PRNG}(x_0)$ and H; recover x_{new} (also derived from x_0); reconstruct $m^* = P(0)$ by Lagrange interpolation through the k points $\{(x_j, s_j)\}_{j=1}^{k-1} \cup \{(x_{\text{new}}, s_{\text{new}})\}$. Success with probability 1 follows from AEAD correctness and uniqueness of the degree-$(k-1)$ interpolant. *(ii) Decoy opening of* m^* *via* dk_0. Enc embeds m^* into c_{FBE} and emits dk_0; Dec on $k = dk_0$ then recovers m^* by FBE correctness [1] (dual-embedding). *(iii) Decoy opening of* m_i*,* $i \geq 1$. Each AddDecoy embeds m_i into the updated c_{FBE} and returns dk_i; Dec on $k = dk_i$ decapsulates, decrypts C_2, and runs FBE decryption, recovering m_i with probability 1 [1]. Distinctness of the $\ell + 1$ messages is enforced by the caller at encryption time. □

B Multi-Message Security

We formalize the seed-reuse attack forward-referenced in Section 6.4 and show that seed ratcheting mitigates it.

Lemma 7 (Key-reuse vulnerability). *If IE encrypts distinct messages* m_1, m_2 *with the same* (x_0, T, k)*, then* $m_1 - m_2 \pmod p$ *is recoverable from* $s_{\text{new},1}$ *and* $s_{\text{new},2}$.

Proof. Both polynomials share the $k-1$ interior points (x_j, s_j). Thus $D(x) := P_1(x) - P_2(x)$ has degree $\leq k-1$ and vanishes at $\{x_j\}_{j=1}^{k-1}$, giving $D(x) = c\prod_{j=1}^{k-1}(x - x_j)$. The adversary observes $D(x_{\text{new}}) = s_{\text{new},1} - s_{\text{new},2}$ at the known abscissa x_{new}, determining c, and therefore $D(0) = m_1 - m_2$. □

Seed ratcheting. Set $x_0^{(i+1)} \leftarrow H(x_0^{(i)} \| C^{(i)})$ after each message. In the random-oracle model, each derived seed is indistinguishable from uniform, so shares and abscissae are independent across messages.

Theorem 8 (Seed-ratcheting security). *Under seed ratcheting with random oracle* H*, for any polynomial-length message sequence and any PPT adversary observing all ciphertexts: (i) per-message IE parameters are computationally independent; (ii) per-message ED and CD advantages remain* $\varepsilon_{\text{ED}} + \mathit{negl}(\lambda)$ *and* $\mathit{negl}(\lambda)$ *respectively; (iii) the attack of Lemma 7 is mitigated.*

Proof (Proof sketch). Standard hybrid argument: in hybrid i, replace $x_0^{(i+1)} = H(\cdot)$ with $\tilde{x}_0^{(i+1)} \xleftarrow{\$} \{0,1\}^\lambda$; any distinguisher contradicts the random-oracle assumption. With independent seeds, the shared-point structure exploited by Lemma 7 vanishes, and per-message ED and CD follow from Theorems 3 and 4 applied independently. Cover-text rotation and ephemeral KEM keys offer complementary defence-in-depth. □

Worked example. Take $p = 2^{256} - 2^{32} - 977$ (our implementation) and threshold $k = 3$. Suppose the adversary observes two PhantomCrypt transcripts sent under seed reuse: the same (x_0, T, k) yields the same PRNG-selected word positions $\{i_1, i_2\}$, the same shares $s_j = H(w_{i_j})$, and the same abscissae $\{x_1, x_2, x_{\text{new}}\}$. The adversary has no FBE key and no x_0, but reads $s_{\text{new},1}$ and $s_{\text{new},2}$ from the two sessions. By Lemma 7,

$$m_1 - m_2 \;\equiv\; \frac{(s_{\text{new},1} - s_{\text{new},2}) \cdot x_1\, x_2}{(x_{\text{new}} - x_1)(x_{\text{new}} - x_2)} \pmod{p},$$

which leaks the pairwise difference without ever recovering a single share or abscissa value. Under seed ratcheting, the abscissae x_1, x_2 in the two sessions are drawn from computationally independent pseudorandom sequences, the shared-point identity $P_1(x_j) = P_2(x_j)$ no longer holds, and the above formula becomes an equation in unknowns the adversary cannot solve.

Scope. Lemma 7 concerns adversaries passively observing multiple transmissions; it does not degrade the single-message CD and ED guarantees of Theorems 4 and 3. A full multi-message CPA/CCA treatment capturing active adversaries who adaptively force decoy openings across sessions remains open and is listed among the problems in Section 9.