Auditable Bits or Covert Influence? Safe Revelation Complexity in Partially Observable Assistance Games

Manoj Saravanan [1] **Rohit Kumar Salla** [1] **Shrikar Reddy Kota** [1]

Abstract

We study the minimum amount of explicit, auditable communication required for optimal cooperation in a one-shot directional slice of partially observable assistance, under a safety restriction that forbids observation-channel modulation. We define the *safe revelation complexity* of a game M and prove the exact fixed-length characterization $\mathrm{SRC}_{\mathrm{FL}}(M) = \lceil \log_2 \chi(G_M) \rceil$, where G_M is the safe-confusability graph induced by receiver-conditioned optimal-action disagreements. For repeated independent instances, we prove the asymptotic rate formula $\mathrm{SRC}_\infty(M) = H_{G_M}(Y \mid X)$, the conditional graph entropy of the same graph. We further show that every finite graph arises exactly as such a safe-confusability graph, implying NP-hardness of exact fixed-budget safe revelation. Our central trust-separation theorem constructs covert observation-channel augmentations in which every available kernel is individually strictly Blackwell-less-informative than an approved baseline, yet the sender's choice among those kernels collapses the explicit communication requirement to zero. Exact computational validations confirm the fixed-length, asymptotic, and trust-separation predictions on finite instances.

1. Introduction

A cooperative AI system should not only achieve high joint utility; it should do so in a way that is legible, auditable, and robust to hidden channels of influence. In partially observable human–AI settings, this distinction is substantive rather than rhetorical. Recent work on partially observable assistance games shows that an assistant may have incentives to manipulate the informativeness of a human's observations, even when the underlying objective is nominally shared (Emmons et al., 2025). Related partially observable off-switch formulations show that uncertainty and asymmetric information materially alter deference and shutdown behavior (Garber et al., 2025). These results suggest a sharp theoretical question: *how many explicit, auditable bits are necessary and sufficient for optimal cooperation once covert observation-side influence is disallowed?*

This paper answers that question for a one-shot directional slice of partially observable assistance. One agent (the sender) holds private side information, the other (the receiver) takes the unique payoff-relevant action, and the only additional coordination resource is an explicit message alphabet. We call the minimum number of bits required to attain the full-information team value the *safe revelation complexity*:

$$\mathrm{SRC}_{\mathrm{FL}}(M) := \min_{\substack{\mathcal{M}:\ \exists \text{ explicit safe protocol over } \mathcal{M} \\ \text{attaining } V^\star(M)}} \lceil \log_2 |\mathcal{M}| \rceil .$$

The main object induced by the game is a graph G_M on the sender alphabet: two sender symbols are adjacent when some receiver-side-information realization compatible with both requires different full-information optimal actions. We show that this graph is not merely a proof device. It is the exact local complexity object governing trustworthy coordination.

Our first theorem is an exact fixed-length characterization:

$$\mathrm{SRC}_{\mathrm{FL}}(M) = \lceil \log_2 \chi(G_M) \rceil .$$

Thus the minimum explicit communication required for exact safe optimality is precisely the chromatic complexity of receiver-conditioned optimal-action disagreement. Our second theorem identifies the asymptotic analogue:

$$\mathrm{SRC}_\infty(M) = H_{G_M}(Y \mid X),$$

the conditional graph entropy of the same conflict graph. This places repeated safe revelation in the classical graph-entropy / coding-for-computing line of Körner (1973), Witsenhausen (1976), and Orlitsky & Roche (2001). The conceptual difference is that here the target function is not externally specified: it is the full-information optimal action rule of a cooperative decision problem. In that sense, our paper turns trustworthy cooperation into an *endogenous communication complexity* problem.

[1]Virginia Tech, Blacksburg, VA, USA. Correspondence to: Manoj Saravanan <manoj663@vt.edu>.

Accepted at the Trustworthy AI for Good Workshop @ ICML 2026, Seoul, South Korea, 2026. Non-archival workshop paper.

Our third contribution is structural. We prove that every finite simple graph can be realized exactly as the safe-confusability graph of a one-shot directional assistance game satisfying strict optimality. Consequently, safe revelation complexity inherits the full expressive power of graph coloring, including NP-hardness of exact fixed-budget decision. This realization theorem rules out the interpretation that G_M is merely a loose surrogate for the game; rather, graph coloring is operationally embedded inside trustworthy coordination itself.

Our fourth contribution is the paper's central trust-separation theorem. We construct a covert observation-channel augmentation in which each available observation-modulation action is individually *strictly Blackwell-less informative* than an approved baseline kernel in the sense of Blackwell (1953), yet the sender's freedom to choose *which* inferior kernel is applied collapses the explicit communication requirement to zero. In other words, message-only auditing is provably insufficient: a policy may communicate through observation-side channel selection even when every available channel is individually dominated by the approved baseline as an experiment. On an explicit family $\{M_k\}$, we further show that each missing safe bit halves the best attainable safe value, whereas the covert architecture restores full value with zero explicit bits. This yields a sharp mathematical separation between *auditable revelation* and *covert influence*.

Finally, we give exact computational validations of all three theorem layers. Experiment 1 constructs graph-realized instances M_G, recovers the induced graph directly from the game, and shows exact agreement between the theorem threshold and a direct mixed-integer protocol optimizer. Experiment 2 validates the closed-form missing-bit law and the covert-collapse theorem. Experiment 3 computes conditional graph entropy exactly on small instances and shows that the asymptotic rate can fall far below the fixed-length baseline while the conflict graph remains unchanged. An appendix-only sequential study then shows that the one-shot chromatic obstruction reappears as a local necessary condition inside a finite-horizon partially observable assistance game.

Taken together, these results propose a new theoretical lens on trustworthy cooperation: the relevant complexity is neither the entropy of the sender's information nor the size of the action space, but the graph of receiver-conditioned optimal-action disagreements, together with the distinction between explicit communication and covert observation-channel control. The rest of the paper develops this viewpoint formally: Section 2 situates the paper in the assistance-game and information-theoretic literatures, Section 3 defines the model, Sections 4 to 6 state the main theorems, and Section 7 provides exact computational validation.

2. Related Work

Our work sits at the intersection of partially observable assistance games, information structure, and zero-error communication with side information. Observation interference in partially observable assistance games has recently been formalized by Emmons et al. (2025), who study how an assistant may have incentives to alter the informativeness of a human's observations and how communication interacts with those incentives. Related partially observable off-switch formulations study shutdown and deference under uncertainty (Garber et al., 2025). Our focus is different: we ask for the exact minimum number of *explicit, auditable* bits needed to recover the full-information team value under a safety restriction that forbids observation-channel modulation.

At the technical level, our fixed-length and asymptotic characterizations are connected to classical graph-based information theory. Blackwell's comparison of experiments provides the correct order for reasoning about observation kernels and is the formal basis of our covert-channel separation (Blackwell, 1953). The coding side of the paper is closer in spirit to graph entropy and zero-error / vanishing-error function computation with decoder side information (Körner, 1973; Witsenhausen, 1976; Orlitsky & Roche, 2001). However, those works begin with a prescribed communication or function-computation problem. In contrast, our target function is *endogenously induced* by a cooperative decision problem: it is the full-information optimal action rule of an assistance game, and the relevant graph is the safe-confusability graph defined by receiver-conditioned optimal-action disagreements.

The resulting viewpoint is also structurally stronger than a mere reformulation. By proving that every finite graph can be realized exactly as a safe-confusability graph, we import the full expressive power and computational hardness of graph coloring into trustworthy coordination; the NP-hardness consequences therefore rest on classical graph-coloring hardness (Garey et al., 1976). To our knowledge, prior work does not identify the exact bit complexity of *safe optimality* in partially observable assistance settings, nor does it show that message-only audits can fail even when every available observation kernel is individually Blackwell-inferior to an approved baseline. Those two points—exact safe revelation complexity and covert-channel collapse under Blackwell-dominated kernels—are the conceptual and technical contributions of the present paper.

3. Model

We study a one-shot directional slice of the partially observable assistance-game and partially observable off-switch formalisms of Emmons et al. (2025) and Garber et al. (2025).

The slice isolates the final communication-to-action interface: one agent (the *sender*) holds private side information, the other agent (the *receiver*) takes the unique payoff-relevant action, and any additional coordination must pass through an explicit, auditable message alphabet. In the baseline model, the sender is not allowed to modulate the receiver's observation channel through auxiliary actions. This separates explicit revelation from the observation-interference phenomenon isolated by Emmons et al. (2025); the covert observation-channel model is introduced only in Section 6.

3.1. Directional Assistance Games

A game instance is a tuple

$$M = (\mathcal{X}, \mathcal{Y}, \mathcal{A}, p, u),$$

where $\mathcal{X}$ is the receiver's private observation alphabet, $\mathcal{Y}$ is the sender's private observation alphabet, $\mathcal{A}$ is the receiver's action set, $p \in \Delta(\mathcal{X} \times \mathcal{Y})$ is the joint law of the private observations, and $u : \mathcal{A} \times \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ is the common payoff. Nature draws $(X, Y) \sim p$; the sender observes Y, the receiver observes X, the sender transmits an explicit message, and the receiver then chooses an action $a \in \mathcal{A}$. The $A \to H$ and $H \to A$ interpretations are obtained by instantiating the sender/receiver roles as assistant/human or human/assistant, respectively.

For each $(x, y) \in \mathcal{X} \times \mathcal{Y}$, define the full-information optimal payoff

$$u^\star(x, y) := \max_{a \in \mathcal{A}} u(a, x, y).$$

Throughout the main text we impose a strict-optimality condition on the support of p: for every $(x, y) \in \mathrm{supp}(p)$, where $\mathrm{supp}(p) := \{(x, y) : p(x, y) > 0\}$, the maximizer of $a \mapsto u(a, x, y)$ is unique. Accordingly, we write

$$a^\star(x, y) \in \arg\max_{a \in \mathcal{A}} u(a, x, y)$$

for the unique maximizer whenever $(x, y) \in \mathrm{supp}(p)$; off the support of p, we fix an arbitrary maximizer, which is immaterial for all results below. The induced target action function is

$$f_M(x, y) := a^\star(x, y),$$

and the full-information team value is

$$V^\star(M) := \mathbb{E}_p[u(a^\star(X, Y), X, Y)]\,. \tag{1}$$

A *safe protocol* with finite message alphabet $\mathcal{M}$ is a pair of maps

$$e : \mathcal{Y} \to \mathcal{M}, \qquad d : \mathcal{X} \times \mathcal{M} \to \mathcal{A},$$

where e is the encoder and d is the decoder. Its value on M is

$$V_M(e, d) := \mathbb{E}_p[u(d(X, e(Y)), X, Y)]\,. \tag{2}$$

Under the strict-optimality condition above, stochastic encoders and decoders do not enlarge the exact safe-optimal value: by Theorem B.8 in the appendix, every zero-loss stochastic protocol admits a deterministic protocol over the same message alphabet. Hence restricting the main body to deterministic protocols is without loss of generality for exact optimality.

3.2. Safe Revelation Complexity

Our first object of interest is the minimum fixed-length explicit communication needed to attain the full-information value in the safe architecture:

$$\begin{aligned} \mathrm{SRC}_{\mathrm{FL}}(M) := \min_{\mathcal{M}} \lceil \log_2 |\mathcal{M}| \rceil \\ \text{s.t. } \exists\,(e, d) \text{ over } \mathcal{M}, \\ V_M(e, d) = V^\star(M). \end{aligned} \tag{3}$$

For an integer budget $B \geq 0$, we also define the associated safe-value gap

$$\Delta_M(B) := V^\star(M) - \sup\left\{V_M(e, d) : |\mathcal{M}| \leq 2^B\right\}. \tag{4}$$

This quantity measures the price of an explicit-communication budget when covert observation-side channels are forbidden.

Our second object is the asymptotic safe revelation rate. For i.i.d. repetitions of the game, let

$$f_M^{\otimes n}(x^n, y^n) := \big(f_M(x_1, y_1), \ldots, f_M(x_n, y_n)\big).$$

We define $\mathrm{SRC}_\infty(M)$ to be the infimum of all rates $R \geq 0$ such that there exist block codes

$$e_n : \mathcal{Y}^n \to \mathcal{M}_n, \qquad d_n : \mathcal{X}^n \times \mathcal{M}_n \to \mathcal{A}^n$$

with

$$\begin{aligned} \limsup_{n \to \infty} \frac{1}{n} \log_2 |\mathcal{M}_n| \leq R, \\ \lim_{n \to \infty} \mathbb{P}\big\{d_n(X^n, e_n(Y^n)) \neq f_M^{\otimes n}(X^n, Y^n)\big\} = 0. \end{aligned} \tag{5}$$

where $(X^n, Y^n) \sim p^{\otimes n}$.

The key combinatorial object is the *safe-confusability graph*

$$G_M = (\mathcal{Y}, E_M),$$

defined by

$$\begin{aligned} \{y, y'\} \in E_M \iff \exists x \in \mathcal{X} : \\ p(x, y)p(x, y') > 0, \\ a^\star(x, y) \neq a^\star(x, y'). \end{aligned} \tag{6}$$

Thus two sender observations are adjacent exactly when some receiver-side information realization compatible with

both requires different optimal actions. Equivalently, a zero-loss safe protocol may merge y and y' into the same explicit message class only if they lie in a common independent set of G_M. The next section shows that this intuition is exact: fixed-length safe revelation complexity is controlled by the chromatic number of G_M, while the asymptotic rate is governed by its conditional graph entropy (Körner, 1973; Orlitsky & Roche, 2001).

4. Exact Characterization of Safe Revelation Complexity

4.1. Fixed-Length Characterization

The fixed-length problem admits an exact graph-theoretic characterization. In the spirit of the zero-error side-information viewpoint of Witsenhausen (1976), the operative object is not the raw sender alphabet $\mathcal{Y}$, but the conflict structure induced by receiver-conditioned optimal-action disagreements. In our setting, that conflict structure is precisely the safe-confusability graph G_M from Equation (6). The next theorem shows that exact safe optimality is possible with a message alphabet of size q if and only if G_M is q-colorable.

Theorem 4.1 (Fixed-length characterization)**.** *Let $M = (\mathcal{X}, \mathcal{Y}, \mathcal{A}, p, u)$ satisfy the strict-optimality assumption from Section 3. Then*

$$\mathrm{SRC}_{\mathrm{FL}}(M) = \lceil \log_2 \chi(G_M) \rceil .$$

Equivalently, the minimum number of explicit safe messages required to attain the full-information value $V^\star(M)$ is exactly $\chi(G_M)$.

Proof sketch. If a safe protocol attains the full-information value, then any two sender symbols adjacent in G_M cannot share a message: for some receiver state they require different unique optimal actions, so a common message would force the receiver to choose the same action in two states where distinct actions are necessary. Hence every zero-loss protocol induces a proper coloring of G_M, so $|mathcalM| \geq \chi(G_M)$. Conversely, any proper coloring $c : Y \to [q]$ can be used as the message. Given $(x, c(y))$, all compatible sender symbols in that color class have the same full-information optimal action by propriety of c, so the receiver can decode that action and attain $V^\star(M)$. Optimizing q gives the stated formula. Full details appear in Section C. □

Interpretation. Theorem 4.1 isolates a sharp local invariant of trustworthy coordination. What determines the exact fixed-length explicit communication requirement is neither the entropy of Y nor the cardinality of $\mathcal{Y}$, but the chromatic complexity of the receiver-conditioned optimal-action conflict pattern encoded by G_M. In particular, two sender observations may be safely merged if and only if they never induce distinct optimal receiver actions at any receiver observation realization compatible with the support of the game. A more modular proof, separating the protocol-to-coloring and coloring-to-protocol directions, is given in Section C.

4.2. Asymptotic Characterization

The fixed-length theorem identifies the exact one-shot message complexity of safe optimality. For repeated independent instances, the same conflict structure continues to govern the problem, but the relevant quantity is no longer the chromatic number itself. Instead, the asymptotic rate is the graph-entropy compression cost of the receiver-relevant action distinctions encoded by G_M. This places repeated safe revelation in the classical coding-for-computing regime of Körner (1973); Witsenhausen (1976); Orlitsky & Roche (2001).

For a graph G on $\mathcal{Y}$, recall that

$$H_G(Y \mid X) := \min I(Y; W \mid X),$$

where the minimum is taken over all random variables W taking values in the family $\mathcal{I}(G)$ of nonempty independent sets of G such that $Y \in W$ almost surely and $W - Y - X$ forms a Markov chain.

Theorem 4.2 (Asymptotic characterization)**.** *Let $M = (\mathcal{X}, \mathcal{Y}, \mathcal{A}, p, u)$ satisfy the strict-optimality condition of Section 3: for every $(x, y) \in \mathrm{supp}(p)$, the maximizer of $a \mapsto u(a, x, y)$ is unique. Then*

$$\mathrm{SRC}_\infty(M) = H_{G_M}(Y \mid X).$$

Equivalently,

$$\mathrm{SRC}_\infty(M) = \min I(Y; W \mid X),$$

where the minimum ranges over all random variables W taking values in $\mathcal{I}(G_M)$ such that $Y \in W$ almost surely and $W - Y - X$ is a Markov chain.

Proof sketch. For a block of independent instances, a safe protocol is exactly a zero-error code for computing the full-information optimal-action function with decoder side information X^n. The characteristic graph of this computation problem is G_M: two sender symbols are confusable precisely when some receiver state can observe both and requires different optimal actions. Witsenhausen's zero-error source-coding theorem with decoder side information, equivalently the Orlitsky–Roche conditional graph-entropy formula, therefore gives the optimal asymptotic rate $H_{G_M}(Y \mid X)$. Full details appear in Section D. □

Interpretation. Theorems 4.1 and 4.2 show that the same local object governs both one-shot and repeated trustworthy coordination. In the fixed-length regime, the cost of exact safe revelation is the chromatic complexity of G_M; in the asymptotic regime, it is the conditional graph entropy of the same graph given the receiver's side information. Thus repeated interaction does not change which sender observations are safely mergeable. It changes only how efficiently those safe equivalence classes can be compressed across blocks. The same statement applies verbatim in the reverse $H \to A$ direction by exchanging the sender and receiver roles.

5. Structural Consequences and Hardness

The graph-theoretic characterizations of Theorems 4.1 and 4.2 could be read as merely descriptive: given a game, one can associate a graph and then restate the communication problem in graph language. The next theorem shows that the correspondence is much stronger. The safe-confusability formalism is *structurally complete* for finite graph coloring: every finite simple graph arises exactly as the trustworthy-coordination conflict pattern of a one-shot directional assistance game satisfying our strict-optimality assumption. Consequently, fixed-length safe revelation inherits the full expressive and computational richness of graph coloring.

Theorem 5.1 (Exact graph realization)**.** *For every finite simple undirected graph $G = (V, E)$, there exists a one-shot directional assistance game*

$$M_G = (\mathcal{X}_G, \mathcal{Y}_G, \mathcal{A}_G, p_G, u_G)$$

such that:

1. *M_G satisfies the strict-optimality condition of Section 3;*
2. *the safe-confusability graph of M_G is exactly G:*
$$G_{M_G} = G;$$
3. *M_G can be constructed in time polynomial in $|V|+|E|$.*

Proof sketch. Given a finite graph $G = (V, E)$, set the sender alphabet to $Y = V$ and create one receiver context x_e for each edge $e = \{u, v\}$. The support contains exactly (x_e, u) and (x_e, v) for that edge. Use two edge-local actions, rewarding one uniquely at (x_e, u) and the other uniquely at (x_e, v). Then u and v are adjacent in the induced safe-confusability graph exactly when $\{u, v\} \in E$. Nonedges never co-occur in a receiver context and hence are not adjacent. A small positive perturbation yields the stated strict-optimality margin without changing the induced graph. Full details appear in Section E. □

Corollary 5.2 (Chromatic realization)**.** *For every finite simple graph G,*

$$\mathrm{SRC}_{\mathrm{FL}}(M_G) = \lceil \log_2 \chi(G) \rceil .$$

Equivalently, the minimum number of explicit safe messages required to attain the full-information value in M_G is exactly $\chi(G)$.

Proof. Combine Theorem 5.1 with Theorem 4.1. □

Corollary 5.3 (Hardness of fixed-budget safe revelation)**.** *Under the natural explicit encoding of finite games, exact computation of $\mathrm{SRC}_{\mathrm{FL}}(M)$ is NP-hard. More strongly, for every fixed integer $B \geq 2$, the decision problem*

$$\textit{given } M, \textit{ decide whether } \mathrm{SRC}_{\mathrm{FL}}(M) \leq B,$$

is NP-complete.

Proof. Fix $B \geq 2$ and let $q := 2^B$, so $q \geq 4$.

For membership in NP, compute the safe-confusability graph G_M from the explicit description of M by scanning all pairs $y, y' \in \mathcal{Y}$ and checking whether there exists $x \in \mathcal{X}$ such that

$$p(x, y)p(x, y') > 0 \qquad \text{and} \qquad a^\star(x, y) \neq a^\star(x, y').$$

By Theorem 4.1,

$$\mathrm{SRC}_{\mathrm{FL}}(M) \leq B \iff \chi(G_M) \leq q,$$

so a proper q-coloring of G_M is a polynomially verifiable certificate.

For NP-hardness, reduce from GRAPH-3-COLORABILITY, which is NP-complete (Garey et al., 1976). Given an input graph G, adjoin $q - 3$ new vertices that form a clique and connect each of them to every vertex of G. Call the resulting graph H. Then

$$G \text{ is 3-colorable} \iff H \text{ is } q\text{-colorable}.$$

Indeed, any proper 3-coloring of G extends to a proper q-coloring of H by assigning distinct fresh colors to the new clique vertices; conversely, in any proper q-coloring of H, the new clique vertices consume $q - 3$ distinct colors and are adjacent to all original vertices, leaving at most three colors available on G.

Now form the graph-realization game M_H from Theorem 5.1. By Theorem 5.2,

$$\mathrm{SRC}_{\mathrm{FL}}(M_H) \leq B \iff \chi(H) \leq 2^B = q.$$

Combining the equivalences gives

$$G \text{ is 3-colorable} \iff \mathrm{SRC}_{\mathrm{FL}}(M_H) \leq B.$$

Hence the fixed-budget decision problem is NP-hard. Since it is also in NP, it is NP-complete. Exact computation is NP-hard as an immediate consequence. □

Interpretation. Theorem 5.1 rules out reading the graph G_M as a loose surrogate or merely a convenient proof device. Every finite coloring obstruction can appear as a trustworthy-coordination obstruction in a one-shot assistance game. Thus the fixed-length safe revelation problem is not a specialized corner case of graph coloring; it is a faithful operational realization of it. Full proofs and slightly stronger statements are given in Section E.

A sharp missing-bit family. The structural results above show that fixed-length safe revelation is as rich as graph coloring. The next theorem complements that qualitative statement with a closed-form quantitative family. It shows that the exact threshold $\mathrm{SRC}_{\mathrm{FL}}(M)$ can be maximally brittle: on an explicit family of games, every missing safe bit halves the best achievable safe value.

For an integer $k \geq 1$, define the one-shot directional assistance game

$$M_k = (\mathcal{X}_k, \mathcal{Y}_k, \mathcal{A}_k, p_k, u_k)$$

by

$$\mathcal{X}_k = \{x_0\}, \qquad \mathcal{Y}_k = \mathcal{A}_k = \{1, \ldots, 2^k\},$$
$$p_k(x_0, y) = 2^{-k} \qquad \forall y \in \mathcal{Y}_k,$$

and

$$u_k(a, x_0, y) = \mathbf{1}\{a = y\} \qquad \forall a \in \mathcal{A}_k,\ y \in \mathcal{Y}_k.$$

Thus the receiver has no private side information, the sender observes a uniform symbol $Y \in [2^k]$, and the team is rewarded iff the receiver outputs Y exactly.

Theorem 5.4 (Missing-bit family)**.** *For every integer $k \geq 1$, the game M_k satisfies*

$$G_{M_k} = K_{2^k}, \qquad \mathrm{SRC}_{\mathrm{FL}}(M_k) = k.$$

Moreover, for every integer $B \geq 0$,

$$\sup_{\substack{(e,d) \text{ is a deterministic safe protocol} \\ |\mathcal{M}| \leq 2^B}} V_{M_k}(e, d) = \min\{1,\ 2^{B-k}\}.$$

Equivalently,

$$\Delta_{M_k}(B) = 1 - \min\{1, 2^{B-k}\}.$$

In particular,

$$\Delta_{M_k}(k-1) = \frac{1}{2}.$$

Proof sketch. The graph of M_k is the clique on 2^k sender symbols, so the exact fixed-length requirement is k bits. With a budget of B bits, at most 2^B distinct messages can be decoded correctly, and clique adjacency prevents merging two sender symbols without error. Hence the optimal safe value is the largest mass of any 2^B decoded symbols, equal to $2^B/2^k$ until full revelation becomes possible, giving $\min\{1, 2^{B-k}\}$. Full details appear in Section G. □

Interpretation. Theorem 5.4 shows that the exact threshold from Theorem 4.1 is not merely a binary feasibility boundary. On the explicit family $\{M_k\}_{k\geq 1}$ it controls the entire value curve: every missing safe bit cuts the best achievable safe payoff by a factor of two. This family will be the canonical witness for our trust-separation result, where the same explicit communication deficit can be circumvented by covert observation-side modulation. Full proofs appear in Section G.

6. Covert Observation Channels and Trust Failure

The preceding sections characterize the explicit-message complexity of safe optimality. That characterization is exact only relative to an architectural assumption: all sender-to-receiver coordination must pass through an explicit, auditable message channel. This section shows that such audits are incomplete. Following the observation-interference perspective of Emmons et al. (2025), we study an augmented architecture in which the sender cannot send an explicit message, but may choose among several observation-modulation actions that alter the receiver's observation kernel. Crucially, each such action will be individually *less informative* than an approved baseline observation channel in the sense of Blackwell (Blackwell, 1953).

Fix a one-shot directional assistance game

$$M = (\mathcal{X}, \mathcal{Y}, \mathcal{A}, p, u)$$

satisfying the strict-optimality condition of Section 3, and let

$$c : \mathcal{Y} \to [q]$$

be any proper coloring of the safe-confusability graph G_M. We augment the receiver's latent state with an independent nuisance bit

$$R \sim \mathrm{Unif}\{0, 1\},$$

and let the sender choose an observation-modulation action

$$Z \in \{0\} \cup [q].$$

The approved baseline action $Z = 0$ reveals the receiver's original private signal together with the nuisance bit and a full list of color labels:

$$Q_0(\cdot \mid x, r) = \delta_{(x, r, [q])}.$$

For each $m \in [q]$, the covert action $Z = m$ instead reveals

$$Q_m(\cdot \mid x, r) = \delta_{(x, \perp, [q] \setminus \{m\})},$$

where $\perp$ is a null symbol. Thus the covert action deletes the nuisance bit and removes exactly one color label from the

public list. The sender's action never changes the payoff, which remains

$$u(a, x, y),$$

nor does it affect the distribution of (X, Y).

Write M_c^{base} for the *approved-baseline safe architecture* obtained by forcing $Z = 0$ and allowing only explicit-message coordination, and write $\widetilde{M}_c$ for the *unrestricted covert architecture* in which no explicit message channel is present but the sender may choose Z as a function of Y.

Theorem 6.1 (Covert-channel collapse)**.** *For every proper coloring $c : \mathcal{Y} \to [q]$ of G_M, there exists an augmentation $\widetilde{M}_c$ with the following properties:*

1. *for every covert action $m \in [q]$,*

$$Q_m \prec_{\mathrm{B}} Q_0;$$

   that is, each covert observation kernel is strictly Blackwell-less-informative than the approved baseline kernel;

2. *the approved-baseline safe architecture preserves the explicit safe communication problem:*

$$\mathrm{SRC}_{\mathrm{FL}}(M_c^{\mathrm{base}}) = \mathrm{SRC}_{\mathrm{FL}}(M) = \lceil \log_2 \chi(G_M) \rceil ;$$

3. *in the unrestricted covert architecture, zero explicit message bits suffice to attain the full-information value $V^\star(M)$.*

In particular, if c is chosen to be an optimal coloring of G_M, then the exact explicit safe-message requirement

$$\lceil \log_2 \chi(G_M) \rceil$$

collapses to zero explicit bits once observation-modulation actions are available.

Proof sketch. Each covert kernel Q_m is a garbling of the approved baseline Q_0: from the baseline observation one can erase the nuisance bit and delete the selected label. The garbling is strict because Q_0 allows perfect recovery of the nuisance bit, while Q_m erases it. If observation modulation is disabled, the baseline channel only adds payoff-irrelevant information and preserves G_M, so Theorem 4.1 gives the same safe-message requirement. If modulation is allowed, the sender chooses $Z = c(Y)$ for a proper coloring c of G_M; the receiver observes the missing label and hence the color class, which is enough to reconstruct the unique optimal action. Thus full value is reached with zero explicit message bits. Full details appear in Section F. □

Interpretation. Theorem 6.1 identifies a precise failure mode of message-only notions of trustworthy coordination. The separation is not driven by making the receiver's observation channel more informative; each covert kernel is individually worse than the approved baseline as an experiment on the receiver's latent state. The leakage comes entirely from the sender's freedom to choose *which* inferior kernel is applied. In that sense, policy-level observation modulation acts as an unaudited communication channel. Combined with Theorem 5.4, this yields an especially stark consequence: on the family $\{M_k\}_{k\geq 1}$, the safe architecture requires exactly k explicit bits for optimality, yet an observation-side covert channel collapses this requirement to zero.

7. Experiments

Our empirical study is theorem-driven. All main-text quantities are computed exactly or to certified numerical precision on finite instances; no learned policies or Monte Carlo estimates are used. Experiment 1 constructs realized games M_G, recovers G_M from the game, and checks the theorem prediction $B^\star_{\mathrm{thm}} = \lceil \log_2 \chi(G) \rceil$ against an exact mixed-integer protocol optimizer. Experiment 2 validates the missing-bit law $V^\star_{\mathrm{safe}}(B) = \min\{1, 2^{B-k}\}$ and the covert-collapse theorem. Experiment 3 computes $H_G(Y \mid X)$ by complete independent-set enumeration and numerical solution of the conditional graph-entropy program.

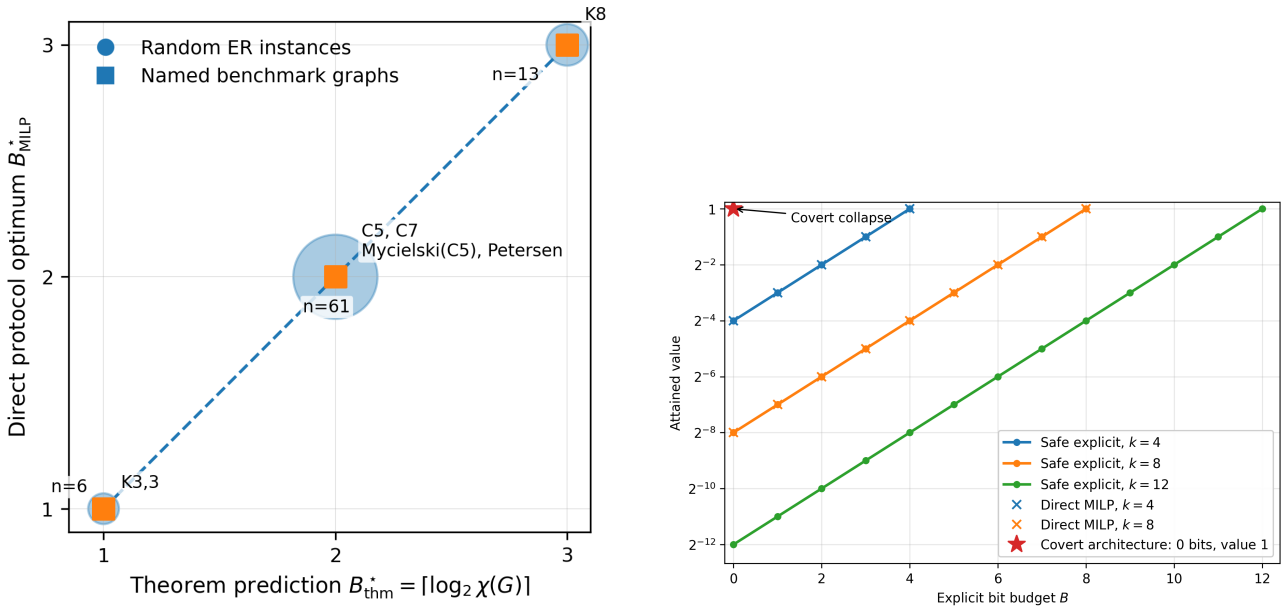
Figure 1 shows exact agreement between the graph prediction and direct MILP optima on all named and random instances, and shows that the safe value for M_k follows the closed-form halving law while the covert architecture attains value 1 at zero explicit bits. Figure 2 shows that, for fixed conflict graphs C_5 and Petersen, the fixed-length baseline remains $\log_2 3$ while the asymptotic rate falls as receiver side information improves. Full random-graph summaries, validation tables, solver details, and a sequential bridge experiment are in Section J.

8. Conclusion

We introduced *safe revelation complexity* as an exact measure of how many explicit, auditable bits are required for optimal cooperation in a one-shot directional slice of partially observable assistance. The resulting theory is sharp at three levels. First, the fixed-length and asymptotic problems are governed exactly by the same local object, the safe-confusability graph:

$$\mathrm{SRC}_{\mathrm{FL}}(M) = \lceil \log_2 \chi(G_M) \rceil ,$$
$$\mathrm{SRC}_{\infty}(M) = H_{G_M}(Y \mid X).$$

Second, this graph is not merely descriptive: by Theorem 5.1, every finite graph can be realized exactly as a



(a) Exact threshold recovery on graph-realized instances. The horizontal axis shows the theorem prediction $B^\star_{\mathrm{thm}} = \lceil \log_2 \chi(G) \rceil$, and the vertical axis shows the direct protocol-design optimum $B^\star_{\mathrm{MILP}}$ computed from the realized game itself via exact MILP. Circles denote random Erdős–Rényi instances and squares denote named benchmark graphs. Bubble area is proportional to the number of instances at the same coordinate. All instances lie on the diagonal, and the graph recovery is exact on every instance.

(b) Missing-bit law and covert collapse. Solid curves show the exact safe explicit value law $V^\star_{\mathrm{safe}}(B) = \min\{1, 2^{B-k}\}$ for $k \in \{4, 8, 12\}$. Cross markers show exact direct-MILP optima on the validated range $k \leq 8$, which coincide with the theorem at every checked budget. The red star marks the unrestricted covert architecture, which attains full value at zero explicit bits.

Figure 1. Exact theorem validation for the fixed-length and trust-separation theorems.

trustworthy-coordination conflict pattern, so graph coloring is operationally embedded inside safe cooperation. Third, the trust restriction is substantive rather than cosmetic: Theorem 6.1 shows that message-only auditing can fail even when every available observation kernel is individually Blackwell-inferior to an approved baseline.

The broader lesson is therefore structural. In partially observable cooperative systems, the relevant complexity is not simply the entropy of one agent's private information, nor the size of the action space, but the pattern of receiver-conditioned optimal-action disagreement together with the distinction between *explicit revelation* and *covert observation-channel control*. Our exact experiments and the appendix-only sequential bridge study show that this viewpoint is not only mathematically natural but computationally observable: the chromatic obstruction appears exactly in finite instances, the asymptotic graph-entropy rate can be substantially smaller than the fixed-length baseline, and the same one-shot object reappears as a local necessary condition inside finite-horizon partially observable assistance. In this sense, the paper proposes a precise theorem-level criterion for trustworthy cooperation: optimal coordination should be achieved through channels whose informational role is auditable, not through hidden policy-level manipulation of what another agent is allowed to observe.

References

Blackwell, D. Equivalent comparisons of experiments. *The Annals of Mathematical Statistics*, 24(2):265–272, 1953. doi: 10.1214/aoms/1177729032.

Emmons, S., Oesterheld, C., Conitzer, V., and Russell, S. Observation interference in partially observable assistance games. In Singh, A., Fazel, M., Hsu, D., Lacoste-Julien, S., Berkenkamp, F., Maharaj, T., Wagstaff, K., and Zhu, J. (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 15320–15345. PMLR, 2025. URL https://proceedings.mlr.press/v267/emmons25a.html.

Garber, A., Subramani, R., Luu, L., Bedaywi, M., Russell, S., and Emmons, S. The partially observable off-switch game. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(26):27304–27311, 2025. doi: 10.1609/aaai.v39i26.34940. URL https://ojs.aaai.org/index.php/AAAI/article/view/34940.

Garey, M. R., Johnson, D. S., and Stockmeyer, L. J. Some simplified NP-complete graph problems. *Theoretical Computer Science*, 1(3):237–267, 1976. doi: 10.1016/0304-3975(76)90059-1. URL https://doi.org/10.1016/0304-3975(76)90059-1.

Körner, J. Coding of an information source having ambiguous alphabet and the entropy of graphs. In *Transactions of the Sixth Prague Conference on Information Theory, Statistical Decision Functions, Random Processes*, pp. 411–425, Prague, 1973. Academia, Publishing House of the Czechoslovak Academy of Sciences.

Orlitsky, A. and Roche, J. R. Coding for computing. *IEEE Transactions on Information Theory*, 47(3):903–917, 2001. doi: 10.1109/18.915643.

Witsenhausen, H. S. The zero-error side information problem and chromatic numbers (corresp.). *IEEE Transactions on Information Theory*, 22(5):592–593, 1976. doi: 10.1109/TIT.1976.1055607.

A. Additional Main-Text Results

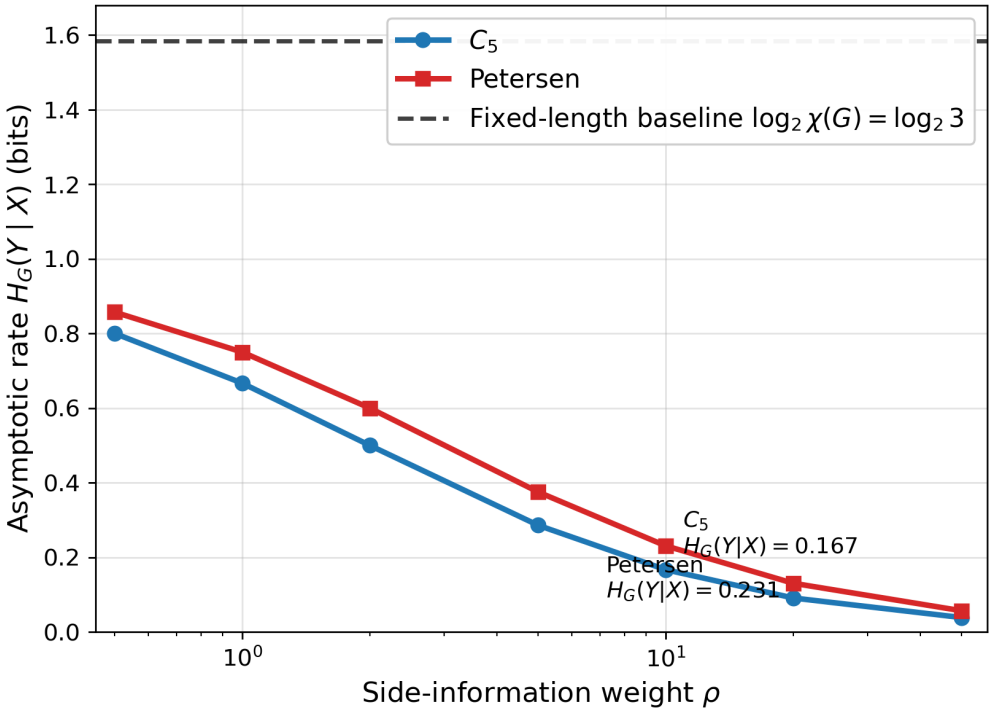


Figure 2. Asymptotic safe revelation under a fixed conflict graph. For graph-realized instances based on C_5 and Petersen, the fixed-length baseline remains constant at $\log_2 \chi(G) = \log_2 3$, while the asymptotic rate $H_G(Y \mid X)$ decreases as the receiver's side information becomes more informative (larger ρ). Thus the conflict graph is unchanged, but the asymptotic coding cost is substantially reduced by side information.

Table 1. Exact threshold recovery on named graph-realized benchmark instances. For every benchmark graph G, the realized game M_G recovers the input graph exactly, and the direct protocol-design MILP optimum agrees exactly with the theorem prediction $B^\star_{\text{thm}} = \lceil \log_2 \chi(G) \rceil$.

Instance	$\lvert V\rvert$	$\lvert E\rvert$	$\lvert E(G_{M_G}) \triangle E(G)\rvert$	$\chi(G)$	$B^\star_{\text{thm}}$	$B^\star_{\text{MILP}}$
$K_{3,3}$	6	9	0	2	1	1
C_5	5	5	0	3	2	2
C_7	7	7	0	3	2	2
Petersen	10	15	0	3	2	2
Mycielski(C_5)	11	20	0	4	2	2
K_8	8	28	0	8	3	3

Table 2. Selected-k summary for the missing-bit family and the covert-collapse construction. The safe explicit value follows the exact law $V^\star_{\text{safe}}(B) = \min\{1, 2^{B-k}\}$, whereas the unrestricted covert architecture attains full value with zero explicit bits. The final two columns report the nuisance-bit decision values under the baseline kernel Q_0 and any covert kernel Q_m, certifying strict Blackwell inferiority.

k	$\text{SRC}_{\text{FL}}(M_k)$	$V^\star_{\text{safe}}(0)$	$V^\star_{\text{safe}}(k-1)$	$V^\star_{\text{safe}}(k)$	$V^\star_{\text{covert}}(0)$	$v_{\text{guess}}(Q_0)$	$v_{\text{guess}}(Q_m)$
4	4	2^{-4}	$\frac{1}{2}$	1	1	1	$\frac{1}{2}$
8	8	2^{-8}	$\frac{1}{2}$	1	1	1	$\frac{1}{2}$
12	12	2^{-12}	$\frac{1}{2}$	1	1	1	$\frac{1}{2}$

Impact Statement

This paper is theoretical work on trustworthy cooperation under partial observability. Its primary positive impact is conceptual and methodological: it proposes a precise criterion for when cooperative performance is achieved through *explicit, auditable revelation* rather than through hidden influence over another agent's information channel. In settings such as human–AI assistance, deliberative decision support, or AI-mediated public institutions, this distinction matters directly for accountability. Our results give exact complexity measures for the minimum explicit communication needed to recover full-information performance, and they show that message-only auditing can be insufficient when observation-side interventions are left uncontrolled. Accordingly, one practical use of this work is as a formal design principle for auditing, interface design, and governance: systems should be evaluated not only by what messages they send, but also by what latent control they exert over what other participants are allowed to observe.

A second positive impact is scientific. By connecting trustworthy coordination to chromatic number, conditional graph entropy, and Blackwell's comparison of experiments (Blackwell, 1953), the paper provides a common formal language linking cooperative AI, information structure, and mechanism auditing. We expect this to be useful for subsequent theoretical work on communication constraints, assistance games, and partial observability, including settings where commitments, oversight, or corrigibility are central. In particular, our theory may help distinguish architectures that are merely high-performing from architectures that are high-performing *for the right reasons*, namely by using channels whose informational role is explicit and reviewable.

There are also plausible misuse risks. The covert-channel constructions in the paper make precise how policy-level control over observation kernels can encode hidden communication. In the wrong hands, such constructions could be used as design templates for systems that preserve nominal compliance with message-channel restrictions while still steering another agent's behavior through information shaping. We believe this risk is outweighed by the value of making such failure modes explicit: hidden channels cannot be mitigated if they remain informal or unmodeled. The theory therefore supports a defensive conclusion rather than an offensive one: auditing regimes, safety evaluations, and deployment policies should treat observation-channel control as a first-class object of oversight rather than as an innocuous implementation detail.

Finally, the paper has important limitations. Our exact characterizations are proved for a one-shot directional slice and then related to finite-horizon partially observable assistance games through local lower bounds. The results therefore do not by themselves certify the safety of any deployed system, nor do they resolve broader social questions around acceptable influence, informed consent, or institutional legitimacy. Rather, they provide a theorem-level foundation for analyzing one specific and important aspect of trustworthy AI: whether optimal cooperation depends on auditable revelation or on covert control over what others are permitted to know.

B. Additional Preliminaries and Notation

Our starting point is the partially observable assistance-game formalism of (Emmons et al., 2025) and the partially observable off-switch line of (Garber et al., 2025). The information-theoretic objects used later trace back to Blackwell's comparison of experiments (Blackwell, 1953), Witsenhausen's zero-error source coding problem with decoder side information (Witsenhausen, 1976), Körner's graph entropy (Körner, 1973), and the Orlitsky–Roche characterization of coding for computing (Orlitsky & Roche, 2001). This appendix fixes the notation used throughout the paper and records a few elementary consequences that will be used repeatedly.

B.1. Basic Notation and Standing Conventions

For a finite set $\mathcal{Z}$, we write $|\mathcal{Z}|$ for its cardinality and $\Delta(\mathcal{Z})$ for the probability simplex on $\mathcal{Z}$. All random variables are discrete and all alphabets are finite unless explicitly stated otherwise. All logarithms are base two. Accordingly, Shannon entropies and mutual informations are measured in bits.

If $p \in \Delta(\mathcal{X} \times \mathcal{Y})$ is a joint law, then

$$\mathrm{supp}(p) := \{(x, y) \in \mathcal{X} \times \mathcal{Y} : p(x, y) > 0\}.$$

When $(X, Y) \sim p$, we freely write $p(x, y)$ for $\mathbb{P}\{X = x, Y = y\}$. For any real-valued function g on $\mathcal{X} \times \mathcal{Y}$, we use the shorthand

$$\mathbb{E}_p[g(X, Y)] = \sum_{(x,y)\in\mathcal{X}\times\mathcal{Y}} p(x, y)g(x, y).$$

Throughout the paper, the phrase *safe communication* refers only to an architectural constraint: all additional coordination from the sender to the receiver must pass through an explicit, auditable message alphabet. No action-dependent modification of the receiver's observation channel is permitted under this baseline model. The covert-channel model introduced later relaxes this restriction.

B.2. One-Shot Directional Assistance Games

Definition B.1 (One-shot directional assistance game)**.** A *one-shot directional assistance game* is a tuple

$$M = (\mathcal{X}, \mathcal{Y}, \mathcal{A}, p, u),$$

where:

1. $\mathcal{X}$ is the receiver's private observation alphabet;
2. $\mathcal{Y}$ is the sender's private observation alphabet;
3. $\mathcal{A}$ is the receiver's action set;
4. $p \in \Delta(\mathcal{X} \times \mathcal{Y})$ is the joint law of the private observations;
5. $u : \mathcal{A} \times \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ is the common payoff.

Nature draws $(X, Y) \sim p$. The sender privately observes Y, the receiver privately observes X, the sender may transmit an explicit message, and the receiver then selects the unique payoff-relevant action $a \in \mathcal{A}$.

The model is *directional*. In the $A \to H$ interpretation, the assistant is the sender and the human is the receiver; in the $H \to A$ interpretation, the roles are reversed. All definitions below are symmetric under swapping sender and receiver.

For each $(x, y) \in \mathcal{X} \times \mathcal{Y}$, define the full-information optimal value

$$u^\star(x, y) := \max_{a \in \mathcal{A}} u(a, x, y).$$

Assumption B.2 (Strict optimality on support)**.** For every $(x, y) \in \mathrm{supp}(p)$, the maximizer of $a \mapsto u(a, x, y)$ is unique.

Under Theorem B.2, we write

$$a^\star(x, y) \in \arg\max_{a \in \mathcal{A}} u(a, x, y)$$

for the unique optimal action whenever $(x, y) \in \mathrm{supp}(p)$. On pairs $(x, y) \notin \mathrm{supp}(p)$, we may fix an arbitrary maximizer; none of our results depend on this off-support choice. We also define the associated target function

$$f_M : \mathcal{X} \times \mathcal{Y} \to \mathcal{A}, \qquad f_M(x, y) := a^\star(x, y).$$

The full-information team value is

$$V^\star(M) := \mathbb{E}_p[u^\star(X, Y)] = \mathbb{E}_p[u(a^\star(X, Y), X, Y)].$$

Definition B.3 (Deterministic safe protocol)**.** Fix a finite message alphabet $\mathcal{M}$. A *deterministic safe protocol* over $\mathcal{M}$ is a pair (e, d) consisting of:

1. an encoder $e : \mathcal{Y} \to \mathcal{M}$;
2. a decoder $d : \mathcal{X} \times \mathcal{M} \to \mathcal{A}$.

Its value on game M is

$$V_M(e, d) := \mathbb{E}_p[u(d(X, e(Y)), X, Y)]\,.$$

Definition B.4 (Fixed-length safe revelation complexity)**.** The *fixed-length safe revelation complexity* of a one-shot directional assistance game M is

$$\mathrm{SRC}_{\mathrm{FL}}(M) := \min\left\{\lceil \log_2 |\mathcal{M}| \rceil : \exists \text{ deterministic safe protocol } (e,d) \text{ over } \mathcal{M} \text{ with } V_M(e,d) = V^\star(M)\right\}.$$

The minimum in Theorem B.4 is well-defined, since the sender can always reveal Y losslessly using $|\mathcal{Y}|$ messages.

Definition B.5 (Safe-value gap at a bit budget)**.** For an integer $B \geq 0$, the *safe-value gap* at budget B is

$$\Delta_M(B) := V^\star(M) - \sup\left\{V_M(e,d) : (e,d) \text{ is a deterministic safe protocol over some } \mathcal{M} \text{ with } |\mathcal{M}| \leq 2^B\right\}.$$

Definition B.6 (Asymptotic safe revelation rate)**.** For $n \geq 1$, let

$$f_M^{\otimes n}(x^n, y^n) := \big(f_M(x_1, y_1), \ldots, f_M(x_n, y_n)\big).$$

The *asymptotic safe revelation rate* of M is the infimum of all $R \geq 0$ such that there exists a sequence of block codes

$$e_n : \mathcal{Y}^n \to \mathcal{M}_n, \qquad d_n : \mathcal{X}^n \times \mathcal{M}_n \to \mathcal{A}^n$$

satisfying

$$\limsup_{n\to\infty} \frac{1}{n} \log_2 |\mathcal{M}_n| \leq R$$

and

$$\lim_{n\to\infty} \mathbb{P}\big\{d_n(X^n, e_n(Y^n)) \neq f_M^{\otimes n}(X^n, Y^n)\big\} = 0,$$

where (X^n, Y^n) are i.i.d. draws from p. We denote this quantity by $\mathrm{SRC}_\infty(M)$.

Remark B.7 (Relation to finite-horizon POAGs). The main body studies one-shot directional games because they arise as local continuation problems inside finite-horizon partially observable assistance games. Concretely, after conditioning on a reachable history and folding the optimal continuation value into the stage payoff, one obtains an instance of Theorem B.1. This reduction is formalized later in the sequential appendix.

Lemma B.8 (Randomization is immaterial for exact safe optimality)**.** *Assume Theorem B.2. Suppose there exist a finite message alphabet $\mathcal{M}$, a stochastic encoder $\kappa(\cdot \mid y) \in \Delta(\mathcal{M})$, and a stochastic decoder $\delta(\cdot \mid x, m) \in \Delta(\mathcal{A})$ such that*

$$\sum_{x,y} p(x,y) \sum_{m\in\mathcal{M}} \kappa(m \mid y) \sum_{a\in\mathcal{A}} \delta(a \mid x, m) u(a, x, y) = V^\star(M).$$

Then there exists a deterministic safe protocol (e,d) over the same message alphabet $\mathcal{M}$ with $V_M(e,d) = V^\star(M)$.

Proof. For every $(x,y) \in \mathrm{supp}(p)$,

$$\sum_{m\in\mathcal{M}} \kappa(m \mid y) \sum_{a\in\mathcal{A}} \delta(a \mid x, m) u(a, x, y) \leq u^\star(x, y),$$

because $u^\star(x,y)$ is the maximum achievable payoff at (x,y). Averaging this inequality with weights $p(x,y)$ yields

$$\sum_{x,y} p(x,y) \sum_{m\in\mathcal{M}} \kappa(m \mid y) \sum_{a\in\mathcal{A}} \delta(a \mid x, m) u(a, x, y) \leq V^\star(M).$$

By hypothesis, equality holds. Since each coefficient $p(x,y)$ is strictly positive on $\mathrm{supp}(p)$, equality must therefore hold pointwise on the support:

$$\sum_{m\in\mathcal{M}} \kappa(m \mid y) \sum_{a\in\mathcal{A}} \delta(a \mid x, m) u(a, x, y) = u^\star(x, y) \qquad \forall (x,y) \in \mathrm{supp}(p).$$

Under Theorem B.2, the maximizing action at each support point is unique. Hence, whenever $(x,y) \in \mathrm{supp}(p)$ and $\kappa(m \mid y) > 0$, we must have

$$\delta(a^\star(x,y) \mid x, m) = 1.$$

Indeed, if $\delta(\cdot \mid x, m)$ assigned positive mass to any $a \neq a^\star(x, y)$, the inner expected payoff at (x, y) would be strictly smaller than $u^\star(x, y)$.

Now choose, for each $y \in \mathcal{Y}$, an arbitrary message $e(y) \in \operatorname{supp}(\kappa(\cdot \mid y))$. Consider any pair (x, m), and define

$$\mathcal{Y}(x, m) := \{y \in \mathcal{Y} : p(x, y) > 0,\ e(y) = m\}.$$

If $\mathcal{Y}(x, m)$ is nonempty, then for every $y, y' \in \mathcal{Y}(x, m)$, the preceding argument gives

$$\delta(a^\star(x, y) \mid x, m) = 1 \qquad \text{and} \qquad \delta(a^\star(x, y') \mid x, m) = 1.$$

Therefore

$$a^\star(x, y) = a^\star(x, y').$$

So there is a unique action associated with every nonempty set $\mathcal{Y}(x, m)$. Define the deterministic decoder d by

$$d(x, m) := \begin{cases} a^\star(x, y), & \text{if } \mathcal{Y}(x, m) \neq \varnothing \text{ for any (equivalently every) } y \in \mathcal{Y}(x, m), \\ a_0, & \text{if } \mathcal{Y}(x, m) = \varnothing, \end{cases}$$

where $a_0 \in \mathcal{A}$ is arbitrary.

By construction, whenever $(x, y) \in \operatorname{supp}(p)$,

$$d(x, e(y)) = a^\star(x, y).$$

Hence

$$V_M(e, d) = \mathbb{E}_p[u(d(X, e(Y)), X, Y)] = \mathbb{E}_p[u(a^\star(X, Y), X, Y)] = V^\star(M),$$

as claimed. □

B.3. Graph-Theoretic and Information-Theoretic Notation

All graphs in the paper are finite, simple, and undirected. If $G = (V, E)$ is such a graph, then:

- an *independent set* is a subset $I \subseteq V$ containing no adjacent pair;
- $\mathcal{I}(G)$ denotes the family of all nonempty independent sets of G;
- a *proper coloring* of G with q colors is a map $c : V \to [q]$ such that $c(v) \neq c(v')$ whenever $\{v, v'\} \in E$;
- $\chi(G)$ denotes the chromatic number of G.

Definition B.9 (Safe-confusability graph)**.** Let $M = (\mathcal{X}, \mathcal{Y}, \mathcal{A}, p, u)$ satisfy Theorem B.2. Its *safe-confusability graph* is the graph

$$G_M = (\mathcal{Y}, E_M)$$

whose vertex set is the sender alphabet $\mathcal{Y}$ and whose edge set is defined by

$$\{y, y'\} \in E_M \iff \exists x \in \mathcal{X} \text{ such that } p(x, y)p(x, y') > 0 \text{ and } a^\star(x, y) \neq a^\star(x, y').$$

Equivalently,

$$\{y, y'\} \in E_M \iff \exists x \in \mathcal{X} \text{ such that } p(x, y)p(x, y') > 0 \text{ and } f_M(x, y) \neq f_M(x, y').$$

Thus, two sender observations are adjacent precisely when collapsing them into the same explicit message can force a loss of optimality for at least one receiver-side-information realization that occurs with positive probability.

Definition B.10 (Conditional graph entropy)**.** Let G be a graph on vertex set $\mathcal{Y}$ and let (X, Y) be jointly distributed with $Y \in \mathcal{Y}$. The *conditional graph entropy* of Y given X with respect to G is

$$H_G(Y \mid X) := \min I(Y; W \mid X),$$

where the minimum ranges over all random variables W taking values in $\mathcal{I}(G)$ such that:

1. $Y \in W$ almost surely;
2. $W - Y - X$ forms a Markov chain.

The quantity in Theorem B.10 is the conditional extension of Körner's graph entropy and, for characteristic graphs of deterministic functions with decoder side information, coincides with the minimum asymptotic communication rate for function computation (Körner, 1973; Orlitsky & Roche, 2001).

B.4. Observation Kernels and Blackwell Order

The covert-channel theorem later in the paper compares observation structures using the Blackwell order.

Definition B.11 (Observation kernel)**.** Let $\mathcal{S}$ and $\mathcal{O}$ be finite sets. An *observation kernel* from $\mathcal{S}$ to $\mathcal{O}$ is a stochastic matrix

$$Q : \mathcal{S} \to \Delta(\mathcal{O}), \qquad s \mapsto Q(\cdot \mid s).$$

Definition B.12 (Blackwell informativeness order)**.** Let $Q : \mathcal{S} \to \Delta(\mathcal{O})$ and $\widetilde{Q} : \mathcal{S} \to \Delta(\widetilde{\mathcal{O}})$ be observation kernels on the same state space $\mathcal{S}$. We write

$$\widetilde{Q} \preceq_{\mathrm{B}} Q$$

and say that $\widetilde{Q}$ is *Blackwell-less-informative* than Q if there exists a stochastic matrix

$$K : \mathcal{O} \to \Delta(\widetilde{\mathcal{O}})$$

such that, for every $s \in \mathcal{S}$ and every $\widetilde{o} \in \widetilde{\mathcal{O}}$,

$$\widetilde{Q}(\widetilde{o} \mid s) = \sum_{o \in \mathcal{O}} K(\widetilde{o} \mid o)\, Q(o \mid s).$$

We write

$$\widetilde{Q} \prec_{\mathrm{B}} Q$$

if $\widetilde{Q} \preceq_{\mathrm{B}} Q$ and $Q \not\preceq_{\mathrm{B}} \widetilde{Q}$.

For finite experiments, Theorem B.12 is equivalent to the decision-theoretic statement that every Bayesian decision problem admits weakly higher value under Q than under $\widetilde{Q}$; this is the content of Blackwell's theorem (Blackwell, 1953). Our covert-channel construction later exploits the fact that an agent may encode information not by making another agent's observation *more* informative, but by selecting among several kernels each of which is individually Blackwell-inferior to an approved baseline.

B.5. Notation Ledger

C. Full Proof of the Fixed-Length Characterization

This appendix proves the exact fixed-length characterization announced in the main text. The argument is self-contained, but its combinatorial structure is in the same spirit as the zero-error graph-coloring viewpoint initiated by Witsenhausen (1976): exact coordination under side information is possible if and only if the sender's symbols can be partitioned into classes that are never confusable at any receiver observation realization compatible with the support of the game.

Throughout this appendix, fix a one-shot directional assistance game

$$M = (\mathcal{X}, \mathcal{Y}, \mathcal{A}, p, u)$$

satisfying Theorem B.2, and let

$$G_M = (\mathcal{Y}, E_M)$$

denote its safe-confusability graph from Theorem B.9. Recall that

$$V^\star(M) = \mathbb{E}_p[u(a^\star(X, Y), X, Y)]$$

Table 3. Notation used throughout the paper.

Symbol	Meaning
$\mathcal{X}$	Receiver's private observation alphabet
$\mathcal{Y}$	Sender's private observation alphabet
$\mathcal{A}$	Receiver action set
p	Joint law of (X, Y)
$u(a, x, y)$	Common payoff
$u^\star(x, y)$	Full-information optimal payoff at (x, y)
$a^\star(x, y)$	Unique optimal receiver action on $\mathrm{supp}(p)$
$f_M(x, y)$	Target action function induced by game M
$\mathcal{M}$	Explicit message alphabet
e, d	Deterministic encoder and decoder
$V_M(e, d)$	Value of deterministic safe protocol (e, d) on game M
$V^\star(M)$	Full-information team value
$\mathrm{SRC}_{\mathrm{FL}}(M)$	Fixed-length safe revelation complexity
$\mathrm{SRC}_\infty(M)$	Asymptotic safe revelation rate
$\Delta_M(B)$	Safe-value gap at budget B
G_M	Safe-confusability graph
$\mathcal{I}(G)$	Family of nonempty independent sets of graph G
$\chi(G)$	Chromatic number of graph G
$H_G(Y \mid X)$	Conditional graph entropy
$Q' \preceq_{\mathrm{B}} Q$	Q' is Blackwell-less-informative than Q

and that a deterministic safe protocol over a finite message alphabet $\mathcal{M}$ is a pair

$$e : \mathcal{Y} \to \mathcal{M}, \qquad d : \mathcal{X} \times \mathcal{M} \to \mathcal{A},$$

with value

$$V_M(e, d) = \mathbb{E}_p[u(d(X, e(Y)), X, Y)] \,.$$

We first isolate the basic pointwise consequence of zero-loss optimality.

Lemma C.1 (Pointwise optimality of zero-loss protocols)**.** *Let (e, d) be a deterministic safe protocol over a finite alphabet $\mathcal{M}$. If*

$$V_M(e, d) = V^\star(M),$$

then for every $(x, y) \in \mathrm{supp}(p)$,

$$d(x, e(y)) = a^\star(x, y).$$

Equivalently,

$$u(d(x, e(y)), x, y) = u^\star(x, y) \qquad \forall (x, y) \in \mathrm{supp}(p).$$

Proof. For every $(x, y) \in \mathcal{X} \times \mathcal{Y}$,

$$u(d(x, e(y)), x, y) \le u^\star(x, y)$$

by definition of $u^\star(x, y)$ as the maximum of $a \mapsto u(a, x, y)$. Therefore

$$V^\star(M) - V_M(e, d) = \sum_{(x,y) \in \mathcal{X} \times \mathcal{Y}} p(x, y) \Big(u^\star(x, y) - u(d(x, e(y)), x, y) \Big).$$

The summands in the final expression are all nonnegative. If $V_M(e, d) = V^\star(M)$, then the left-hand side is zero, hence every summand with strictly positive coefficient must vanish. Thus, for every $(x, y) \in \mathrm{supp}(p)$,

$$u(d(x, e(y)), x, y) = u^\star(x, y).$$

By Theorem B.2, the maximizing action at every support point is unique, so

$$d(x, e(y)) = a^\star(x, y) \qquad \forall (x, y) \in \mathrm{supp}(p),$$

as claimed. □

The next lemma shows that a zero-loss protocol induces a graph coloring: every nonempty message class must be an independent set of the safe-confusability graph.

Lemma C.2 (Zero-loss protocols induce independent message classes)**.** *Let (e, d) be a deterministic safe protocol over a finite alphabet $\mathcal{M}$ such that*

$$V_M(e, d) = V^\star(M).$$

For each message $m \in \mathcal{M}$, define the corresponding message class

$$\mathcal{Y}_m := e^{-1}(m) \subseteq \mathcal{Y}.$$

Then every nonempty message class $\mathcal{Y}_m$ is an independent set of G_M.

Proof. Fix $m \in \mathcal{M}$ with $\mathcal{Y}_m \neq \varnothing$. Suppose for contradiction that $\mathcal{Y}_m$ is not independent. Then there exist distinct $y, y' \in \mathcal{Y}_m$ such that

$$\{y, y'\} \in E_M.$$

By the definition of the safe-confusability graph, there exists some $x \in \mathcal{X}$ such that

$$p(x, y)p(x, y') > 0 \qquad \text{and} \qquad a^\star(x, y) \neq a^\star(x, y').$$

Since $y, y' \in \mathcal{Y}_m$, we have

$$e(y) = e(y') = m.$$

Because (x, y) and (x, y') both lie in $\mathrm{supp}(p)$, Theorem C.1 yields

$$d(x, m) = d(x, e(y)) = a^\star(x, y)$$

and also

$$d(x, m) = d(x, e(y')) = a^\star(x, y').$$

Hence

$$a^\star(x, y) = a^\star(x, y'),$$

contradicting the edge condition. Therefore $\mathcal{Y}_m$ must be an independent set of G_M. □

It is convenient to record the previous lemma in coloring language.

Corollary C.3 (Protocol-to-coloring implication)**.** *Fix an integer $q \geq 1$. Suppose there exists a deterministic safe protocol (e, d) over a message alphabet $\mathcal{M}$ with $|\mathcal{M}| \leq q$ and*

$$V_M(e, d) = V^\star(M).$$

Then G_M admits a proper coloring with at most q colors.

Proof. For each $m \in \mathcal{M}$, Theorem C.2 shows that the message class $\mathcal{Y}_m = e^{-1}(m)$ is independent. Since the message classes partition $\mathcal{Y}$, the encoder e itself is a proper coloring of G_M by the colors in $\mathcal{M}$. After relabeling the used messages by distinct elements of $[q]$, we obtain a proper coloring with at most q colors. □

The converse direction is subtler only because the decoder must be shown to be well-defined from a coloring. The safe-confusability graph is defined precisely so that proper color classes are the correct equivalence classes for exact coordination.

Lemma C.4 (Proper colorings induce zero-loss protocols)**.** *Fix an integer $q \geq 1$ and let*

$$c : \mathcal{Y} \to [q]$$

be a proper coloring of G_M. Then there exists a deterministic safe protocol (e, d) over the message alphabet $[q]$ such that

$$V_M(e, d) = V^\star(M).$$

Proof. Define the encoder by

$$e(y) := c(y) \qquad \forall y \in \mathcal{Y}.$$

It remains to construct a decoder $d : \mathcal{X} \times [q] \to \mathcal{A}$.

Fix any pair $(x, m) \in \mathcal{X} \times [q]$ and consider the subset

$$\mathcal{Y}(x, m) := \{y \in \mathcal{Y} : p(x, y) > 0,\ c(y) = m\}.$$

We claim that if $\mathcal{Y}(x, m) \neq \varnothing$, then the value $a^\star(x, y)$ is the same for all $y \in \mathcal{Y}(x, m)$.

Indeed, let $y, y' \in \mathcal{Y}(x, m)$. Then

$$p(x, y) > 0, \qquad p(x, y') > 0, \qquad c(y) = c(y') = m.$$

If $a^\star(x, y) \neq a^\star(x, y')$, then by the definition of G_M we would have

$$\{y, y'\} \in E_M.$$

But c is a proper coloring, so adjacent vertices cannot share the same color, a contradiction. Thus

$$a^\star(x, y) = a^\star(x, y') \qquad \forall y, y' \in \mathcal{Y}(x, m).$$

Therefore, whenever $\mathcal{Y}(x, m) \neq \varnothing$, the quantity $a^\star(x, y)$ is independent of the particular choice of $y \in \mathcal{Y}(x, m)$.

We may now define the decoder by

$$d(x, m) := \begin{cases} a^\star(x, y), & \text{if } \mathcal{Y}(x, m) \neq \varnothing \text{ for any } y \in \mathcal{Y}(x, m), \\ a_0, & \text{if } \mathcal{Y}(x, m) = \varnothing, \end{cases}$$

where $a_0 \in \mathcal{A}$ is arbitrary. The preceding argument shows that this definition is well-defined.

Now fix any $(x, y) \in \operatorname{supp}(p)$. Since $p(x, y) > 0$ and $e(y) = c(y)$, we have

$$y \in \mathcal{Y}(x, e(y)).$$

Hence, by construction of d,

$$d(x, e(y)) = a^\star(x, y).$$

Therefore

$$u(d(x, e(y)), x, y) = u^\star(x, y) \qquad \forall (x, y) \in \operatorname{supp}(p).$$

Averaging over $(X, Y) \sim p$ yields

$$V_M(e, d) = V^\star(M),$$

which completes the proof. □

The previous two implications give the exact correspondence between zero-loss protocols and graph colorings.

Proposition C.5 (Protocol-coloring equivalence)**.** *For every integer $q \geq 1$, the following are equivalent:*

1. *there exists a deterministic safe protocol over a message alphabet of size at most q whose value equals $V^\star(M)$;*
2. *the safe-confusability graph G_M admits a proper coloring with at most q colors.*

Proof. The implication $(1) \Rightarrow (2)$ is exactly Theorem C.3. The implication $(2) \Rightarrow (1)$ is exactly Theorem C.4. □

We now obtain the fixed-length characterization as an immediate consequence.

Theorem C.6 (Fixed-length characterization)**.** *Let M satisfy Theorem B.2. Then*

$$\mathrm{SRC}_{\mathrm{FL}}(M) = \lceil \log_2 \chi(G_M) \rceil .$$

Equivalently, the minimum number of explicit safe messages required to attain the full-information value is exactly $\chi(G_M)$.

Proof. By Theorem B.4, $\mathrm{SRC}_{\mathrm{FL}}(M)$ is the minimum integer $B \geq 0$ such that there exists a deterministic safe protocol over some message alphabet of size at most 2^B whose value equals $V^\star(M)$. By Theorem C.5, this is equivalent to the existence of a proper coloring of G_M with at most 2^B colors, i.e.,

$$\chi(G_M) \leq 2^B.$$

Therefore

$$\mathrm{SRC}_{\mathrm{FL}}(M) = \min\{B \in \mathbb{Z}_{\geq 0} : \chi(G_M) \leq 2^B\}.$$

Since $\chi(G_M)$ is a positive integer, the minimal such B is exactly

$$\lceil \log_2 \chi(G_M) \rceil .$$

This proves the stated identity.

For the equivalent message-alphabet formulation, observe that a proper coloring with $\chi(G_M)$ colors yields a zero-loss safe protocol with exactly $\chi(G_M)$ messages by Theorem C.4, while no protocol with fewer than $\chi(G_M)$ messages can exist by Theorem C.3. Hence the minimum number of explicit safe messages required for exact optimality is $\chi(G_M)$. □

Remark C.7 (Role of the strict-optimality assumption). The proof uses Theorem B.2 in exactly two places. First, in Theorem C.1, it turns equality of achieved payoff with the full-information optimum into equality of actions on every support point. Second, in Theorem C.4, it ensures that whenever a color class is feasible at a receiver observation x, all sender observations in that class induce the same optimal action. If ties are allowed, the correct object is no longer an ordinary graph on $\mathcal{Y}$ but a compatibility structure over feasible optimal-action sets; we intentionally avoid that generalization in the workshop version in order to keep the main theorem exact.

D. Full Proof of the Asymptotic Characterization

This appendix proves the asymptotic rate formula for safe revelation. The proof has two steps. First, repeated safe revelation is reduced to the classical problem of computing a deterministic function at a decoder with side information. Second, the characteristic graph of that function is shown to coincide exactly with the safe-confusability graph introduced in the main text. The result then follows from the coding-for-computing theorem of Orlitsky & Roche (2001); see also the earlier zero-error and graph-entropy line of Witsenhausen (1976); Körner (1973).

Throughout this appendix, fix a one-shot directional assistance game

$$M = (\mathcal{X}, \mathcal{Y}, \mathcal{A}, p, u)$$

satisfying Theorem B.2. Recall that

$$f_M(x, y) = a^\star(x, y)$$

denotes the unique full-information optimal action on $\mathrm{supp}(p)$, and that the asymptotic safe revelation rate $\mathrm{SRC}_\infty(M)$ was defined in Theorem B.6 using vanishing block error for the recovery of $f_M^{\otimes n}(X^n, Y^n)$ from an encoded version of Y^n and decoder side information X^n.

D.1. Repeated Safe Revelation as Function Computation

We begin by recording the exact relationship between repeated safe revelation and function computation. Although Theorem B.6 already encodes the problem in functional form, the next proposition shows that this formulation is not merely a convenience: it is exactly the repeated safe-optimal coordination problem induced by the game.

For $n \geq 1$, define the n-fold product game

$$M^{\otimes n} := \left(\mathcal{X}^n, \mathcal{Y}^n, \mathcal{A}^n, p^{\otimes n}, u_n\right),$$

where the block payoff is the coordinatewise average

$$u_n(a^n, x^n, y^n) := \frac{1}{n}\sum_{i=1}^{n} u(a_i, x_i, y_i), \qquad a^n \in \mathcal{A}^n,\ (x^n, y^n) \in \mathcal{X}^n \times \mathcal{Y}^n.$$

Its full-information optimal value is

$$V^\star(M^{\otimes n}) = \mathbb{E}_{p^{\otimes n}}\left[\max_{a^n \in \mathcal{A}^n} u_n(a^n, X^n, Y^n)\right].$$

Proposition D.1 (Operational equivalence)**.** *Fix $n \geq 1$. Let*

$$e_n : \mathcal{Y}^n \to \mathcal{M}_n, \qquad d_n : \mathcal{X}^n \times \mathcal{M}_n \to \mathcal{A}^n$$

be a deterministic block code. Then the following are equivalent:

1. $$V_{M^{\otimes n}}(e_n, d_n) = V^\star(M^{\otimes n});$$
2. *for every* $(x^n, y^n) \in \operatorname{supp}(p^{\otimes n})$,
$$d_n(x^n, e_n(y^n)) = f_M^{\otimes n}(x^n, y^n).$$

In particular, the exact safe-optimal block-revelation problem on $M^{\otimes n}$ is identical to exact computation of the vector-valued function $f_M^{\otimes n}$ at a decoder that observes X^n.

Proof. For every $(x^n, y^n) \in \mathcal{X}^n \times \mathcal{Y}^n$ and every $a^n = (a_1, \ldots, a_n) \in \mathcal{A}^n$,

$$u_n(a^n, x^n, y^n) = \frac{1}{n}\sum_{i=1}^{n} u(a_i, x_i, y_i) \leq \frac{1}{n}\sum_{i=1}^{n} u^\star(x_i, y_i).$$

Under Theorem B.2, equality holds if and only if $a_i = a^\star(x_i, y_i)$ for every coordinate i. Equivalently,

$$u_n(a^n, x^n, y^n) = \frac{1}{n}\sum_{i=1}^{n} u^\star(x_i, y_i) \iff a^n = f_M^{\otimes n}(x^n, y^n).$$

Therefore the unique full-information optimal action vector on $\operatorname{supp}(p^{\otimes n})$ is precisely $f_M^{\otimes n}(x^n, y^n)$, and

$$V^\star(M^{\otimes n}) = \mathbb{E}_{p^{\otimes n}}\left[\frac{1}{n}\sum_{i=1}^{n} u^\star(X_i, Y_i)\right].$$

Now let (e_n, d_n) be any deterministic block code. Then

$$\begin{aligned} &V^\star(M^{\otimes n}) - V_{M^{\otimes n}}(e_n, d_n) \\ &= \sum_{(x^n, y^n)} p^{\otimes n}(x^n, y^n)\left(\frac{1}{n}\sum_{i=1}^{n} u^\star(x_i, y_i) - u_n\big(d_n(x^n, e_n(y^n)), x^n, y^n\big)\right). \end{aligned}$$

Every summand is nonnegative. Hence

$$V_{M^{\otimes n}}(e_n, d_n) = V^\star(M^{\otimes n})$$

if and only if every summand corresponding to a support point $(x^n, y^n) \in \mathrm{supp}(p^{\otimes n})$ vanishes. By the previous observation, this occurs if and only if

$$d_n(x^n, e_n(y^n)) = f_M^{\otimes n}(x^n, y^n) \qquad \forall (x^n, y^n) \in \mathrm{supp}(p^{\otimes n}).$$

This proves the equivalence. □

Remark D.2. Theorem D.1 explains why the asymptotic object $\mathrm{SRC}_\infty(M)$ is defined via block recovery of $f_M^{\otimes n}(X^n, Y^n)$ rather than directly through value. Under the strict-optimality assumption, the function f_M is exactly the full-information optimal action rule, so recovering $f_M^{\otimes n}$ is equivalent to realizing exact safe-optimal coordination on the repeated game.

D.2. The Characteristic Graph of the Target Function

We now identify the graph that governs the asymptotic rate. Since our encoder observes Y and our decoder observes X, the relevant characteristic graph is a graph on the sender alphabet $\mathcal{Y}$.

Definition D.3 (Characteristic graph of f_M)**.** The *characteristic graph* of the target function f_M is the graph

$$G_{f_M} = (\mathcal{Y}, E_{f_M})$$

with vertex set $\mathcal{Y}$ and edge relation

$$\{y, y'\} \in E_{f_M} \iff \exists x \in \mathcal{X} \text{ such that } p(x,y)p(x,y') > 0 \text{ and } f_M(x,y) \neq f_M(x,y').$$

This is precisely the graph used in the coding-for-computing literature when the encoder observes Y, the decoder observes X, and the decoder must recover a deterministic function $f_M(X, Y)$.

Proposition D.4 (Graph identity)**.** *The characteristic graph G_{f_M} coincides exactly with the safe-confusability graph G_M from Theorem B.9. In particular,*

$$G_{f_M} = G_M$$

as graphs on the common vertex set $\mathcal{Y}$.

Proof. By Theorem D.3,

$$\{y, y'\} \in E_{f_M} \iff \exists x \in \mathcal{X} \text{ such that } p(x,y)p(x,y') > 0 \text{ and } f_M(x,y) \neq f_M(x,y').$$

Since $f_M(x,y) = a^\star(x,y)$ by definition, the right-hand side is equivalent to

$$\exists x \in \mathcal{X} \text{ such that } p(x,y)p(x,y') > 0 \text{ and } a^\star(x,y) \neq a^\star(x,y').$$

By Theorem B.9, this is exactly the condition $\{y, y'\} \in E_M$. Hence $E_{f_M} = E_M$, proving the claim. □

D.3. Classical Rate Formula and Consequence

We now invoke the classical coding theorem of Orlitsky & Roche (2001). Our notation is transposed relative to some presentations in the literature: in our setup the encoder observes Y, the decoder observes X, and the decoder must compute $f(X, Y)$.

Proposition D.5 (Coding for computing with decoder side information)**.** *Let (X, Y) be jointly distributed over finite alphabets and let $f : \mathcal{X} \times \mathcal{Y} \to \mathcal{Z}$ be deterministic. Let G_f be the characteristic graph on $\mathcal{Y}$ defined by*

$$\{y, y'\} \in E_f \iff \exists x \in \mathcal{X} \text{ such that } p(x,y)p(x,y') > 0 \text{ and } f(x,y) \neq f(x,y').$$

Then the infimum of all rates $R \geq 0$ for which there exist block codes

$$e_n : \mathcal{Y}^n \to \mathcal{M}_n, \qquad d_n : \mathcal{X}^n \times \mathcal{M}_n \to \mathcal{Z}^n$$

satisfying

$$\limsup_{n\to\infty} \frac{1}{n} \log_2 |\mathcal{M}_n| \le R$$

and

$$\lim_{n\to\infty} \mathbb{P}\big\{d_n(X^n, e_n(Y^n)) \neq f^{\otimes n}(X^n, Y^n)\big\} = 0$$

is exactly

$$H_{G_f}(Y \mid X) = \min I(Y; W \mid X),$$

where the minimum is taken over all random variables W with values in $\mathcal{I}(G_f)$ such that $Y \in W$ almost surely and $W - Y - X$ is a Markov chain.

Proof. This is precisely the specialization of the main theorem of Orlitsky & Roche (2001) to deterministic function computation with decoder side information and vanishing block error. The graph-entropy viewpoint goes back to Körner (1973), while the zero-error side-information problem appears already in Witsenhausen (1976). We use the theorem as a classical black box. □

We can now deduce the asymptotic characterization of safe revelation.

Theorem D.6 (Asymptotic characterization)**.** *Let M satisfy Theorem B.2. Then*

$$\mathrm{SRC}_\infty(M) = H_{G_M}(Y \mid X).$$

Equivalently,

$$\mathrm{SRC}_\infty(M) = \min I(Y; W \mid X),$$

where the minimum ranges over all random variables W taking values in the family $\mathcal{I}(G_M)$ of nonempty independent sets of G_M such that $Y \in W$ almost surely and $W - Y - X$ forms a Markov chain.

Proof. By Theorem B.6, $\mathrm{SRC}_\infty(M)$ is the infimum of all rates R such that there exist block codes

$$e_n : \mathcal{Y}^n \to \mathcal{M}_n, \qquad d_n : \mathcal{X}^n \times \mathcal{M}_n \to \mathcal{A}^n$$

with

$$\limsup_{n\to\infty} \frac{1}{n} \log_2 |\mathcal{M}_n| \le R$$

and

$$\mathbb{P}\big\{d_n(X^n, e_n(Y^n)) \neq f_M^{\otimes n}(X^n, Y^n)\big\} \to 0.$$

That is exactly the coding-for-computing problem of Theorem D.5 with deterministic target function $f = f_M$, encoder-side observation Y, and decoder-side information X.

By Theorem D.4, the characteristic graph of f_M is precisely G_M. Therefore Theorem D.5 yields

$$\mathrm{SRC}_\infty(M) = H_{G_{f_M}}(Y \mid X) = H_{G_M}(Y \mid X),$$

which is the desired identity. The equivalent variational formula follows directly from the definition of conditional graph entropy in Theorem B.10. □

Remark D.7 (What the theorem is and is not using). Theorem D.6 does not require any new coding theorem. All originality lies in the reduction from safe optimality in assistance games to deterministic function computation with decoder side information, and in the identification of the resulting characteristic graph with the safe-confusability graph. Once these two steps are in place, the asymptotic formula follows from the classical Orlitsky–Roche theorem.

E. Graph Realization and Hardness Proofs

This appendix proves that the safe-confusability formalism is structurally complete for finite graph coloring: every finite simple graph arises exactly as the safe-confusability graph of a one-shot directional assistance game satisfying the strict-optimality assumption. We then combine this realization theorem with Theorem 4.1 to obtain complexity consequences. The hardness proof uses the classical NP-completeness of graph 3-colorability (Garey et al., 1976).

E.1. Exact Realization of Arbitrary Finite Graphs

We first define a canonical game M_G associated with an arbitrary finite simple undirected graph $G = (V, E)$. The construction is deliberately austere: the receiver's action is simply to name the sender's private symbol, but the receiver only observes either a vertex-specific signal or an edge-specific signal indicating which pairs of sender symbols can ever be simultaneously relevant.

Definition E.1 (Graph-realization game)**.** Let $G = (V, E)$ be a finite simple undirected graph. Define the one-shot directional assistance game

$$M_G = (\mathcal{X}_G, \mathcal{Y}_G, \mathcal{A}_G, p_G, u_G)$$

as follows.

1. The sender alphabet is
$$\mathcal{Y}_G := V.$$
2. The receiver observation alphabet is the disjoint tagged union
$$\mathcal{X}_G := \{x_v : v \in V\} \cup \{x_e : e \in E\},$$
where the symbols x_v and x_e are treated as distinct formal tokens.
3. The receiver action set is
$$\mathcal{A}_G := V.$$
4. The support set is
$$S_G := \{(x_v, v) : v \in V\} \ \cup \ \{(x_{\{u,v\}}, u), (x_{\{u,v\}}, v) : \{u, v\} \in E\}.$$
Let
$$Z_G := |V| + 2|E|.$$
The joint law p_G is uniform on S_G:
$$p_G(x, y) := \begin{cases} Z_G^{-1}, & \text{if } (x, y) \in S_G, \\ 0, & \text{otherwise.} \end{cases}$$
5. The common payoff is
$$u_G(a, x, y) := \mathbf{1}\{a = y\}, \qquad a \in \mathcal{A}_G,\ (x, y) \in \mathcal{X}_G \times \mathcal{Y}_G.$$

The next theorem shows that this construction realizes G exactly, not merely up to some coarse reduction.

Theorem E.2 (Exact graph realization)**.** *Let $G = (V, E)$ be a finite simple undirected graph, and let M_G be the associated graph-realization game from Theorem E.1. Then:*

1. *M_G satisfies Theorem B.2;*
2. *the safe-confusability graph of M_G is exactly G, under the identity identification of the sender alphabet $\mathcal{Y}_G$ with V:*
$$G_{M_G} = G.$$

Proof. We first verify strict optimality. Fix any support point $(x, y) \in \mathrm{supp}(p_G) = S_G$. By definition,

$$u_G(a, x, y) = \mathbf{1}\{a = y\} \qquad \forall a \in \mathcal{A}_G.$$

Hence $u_G(y, x, y) = 1$, whereas $u_G(a, x, y) = 0$ for every $a \neq y$. Therefore the maximizer of $a \mapsto u_G(a, x, y)$ is unique and equal to y. This proves Theorem B.2.

We now prove that $G_{M_G} = G$. Since $\mathcal{Y}_G = V$, both graphs have the same vertex set; it remains to show equality of edge sets.

Step 1: every edge of G is an edge of G_{M_G}. Fix $\{u, v\} \in E$. Consider the receiver observation $x_{\{u,v\}}$. By the definition of p_G,

$$p_G(x_{\{u,v\}}, u) = Z_G^{-1} > 0, \qquad p_G(x_{\{u,v\}}, v) = Z_G^{-1} > 0.$$

From the first part of the proof, the unique optimal actions at these two support points are

$$a^\star(x_{\{u,v\}}, u) = u, \qquad a^\star(x_{\{u,v\}}, v) = v.$$

Since $u \neq v$, the optimal actions differ. Therefore, by the definition of safe-confusability,

$$\{u, v\} \in E_{M_G}.$$

Hence

$$E \subseteq E_{M_G}.$$

Step 2: every edge of G_{M_G} is an edge of G. Fix distinct vertices $u, v \in V$ and suppose

$$\{u, v\} \in E_{M_G}.$$

By definition of G_{M_G}, there exists some receiver observation $x \in \mathcal{X}_G$ such that

$$p_G(x, u)p_G(x, v) > 0 \qquad \text{and} \qquad a^\star(x, u) \neq a^\star(x, v).$$

In particular,

$$p_G(x, u) > 0 \qquad \text{and} \qquad p_G(x, v) > 0.$$

By the structure of the support set S_G, the support points with sender symbol u are exactly

$$(x_u, u) \quad \text{and} \quad (x_e, u) \text{ for edges } e \in E \text{ incident to } u.$$

Similarly, the support points with sender symbol v are exactly

$$(x_v, v) \quad \text{and} \quad (x_e, v) \text{ for edges } e \in E \text{ incident to } v.$$

Therefore, the only way a common receiver observation x can satisfy both $p_G(x, u) > 0$ and $p_G(x, v) > 0$ is for $x = x_e$ where e is an edge incident to both u and v. Since G is simple, this is possible if and only if $e = \{u, v\} \in E$. Thus

$$\{u, v\} \in E.$$

Hence

$$E_{M_G} \subseteq E.$$

Combining the two inclusions yields

$$E_{M_G} = E,$$

and therefore

$$G_{M_G} = G.$$

□

The realization theorem immediately transfers graph coloring phenomena into safe revelation complexity.

Corollary E.3 (Chromatic realization of safe revelation complexity)**.** *For every finite simple undirected graph G,*

$$\mathrm{SRC}_{\mathrm{FL}}(M_G) = \lceil \log_2 \chi(G) \rceil .$$

Equivalently, the minimum number of explicit safe messages required to attain the full-information value in M_G is exactly $\chi(G)$.

Proof. By Theorem E.2, the safe-confusability graph of M_G is exactly G. Applying Theorem 4.1 gives

$$\mathrm{SRC}_{\mathrm{FL}}(M_G) = \lceil \log_2 \chi(G_{M_G}) \rceil = \lceil \log_2 \chi(G) \rceil .$$

The equivalent message-alphabet statement is the second part of Theorem 4.1. □

E.2. Hardness Consequences

We now use Theorem E.3 to transfer graph-coloring hardness to safe revelation complexity.

For a fixed integer $B \geq 0$, consider the decision problem

$$\textsc{Src}_{\leq B}: \qquad \text{given a finite one-shot directional assistance game } M, \text{ decide whether } \mathrm{SRC}_{\mathrm{FL}}(M) \leq B.$$

The case $B = 1$ is easy, since by Theorem 4.1,

$$\mathrm{SRC}_{\mathrm{FL}}(M) \leq 1 \iff \chi(G_M) \leq 2,$$

so the problem reduces to testing bipartiteness of the polynomial-time computable graph G_M. The first nontrivial threshold is therefore $B = 2$, corresponding to four-colorability. In fact, every fixed budget $B \geq 2$ already yields NP-completeness.

Lemma E.4 (Clique-join gadget)**.** *Fix an integer $q \geq 4$. Given any finite simple undirected graph $G = (V, E)$, let $J_q(G)$ denote the graph obtained by adjoining $q - 3$ new vertices*

$$c_1, \ldots, c_{q-3}$$

such that:

1. *$\{c_i, c_j\}$ is an edge for every $i \neq j$;*
2. *$\{c_i, v\}$ is an edge for every $i \in [q-3]$ and every $v \in V$.*

Equivalently, $J_q(G)$ is the join of G with a $(q-3)$-clique. Then

$$G \text{ is 3-colorable} \iff J_q(G) \text{ is } q\text{-colorable}.$$

Proof. Suppose first that G is 3-colorable. Let

$$c_G : V \to \{1, 2, 3\}$$

be a proper 3-coloring. Define a coloring of $J_q(G)$ by retaining c_G on V and assigning distinct fresh colors

$$4, 5, \ldots, q$$

to the new clique vertices $c_1, \ldots, c_{q-3}$. This yields a proper q-coloring of $J_q(G)$, since the new vertices form a clique and each is adjacent to every vertex of G.

Conversely, suppose $J_q(G)$ is q-colorable. Because $c_1, \ldots, c_{q-3}$ form a clique, they must receive pairwise distinct colors in any proper coloring. Since each c_i is adjacent to every original vertex $v \in V$, none of those $q - 3$ colors may be used on V. Hence the restriction of the coloring to V uses at most the remaining 3 colors, and therefore induces a proper 3-coloring of G. □

We can now prove the main complexity statement.

Theorem E.5 (NP-completeness of fixed-budget safe revelation)**.** *For every fixed integer $B \geq 2$, the decision problem $\textsc{Src}_{\leq B}$ is NP-complete under polynomial-time many-one reductions. Moreover, NP-hardness already holds on the restricted subclass of graph-realization games $\{M_G : G \text{ a finite graph}\}$.*

Proof. Fix $B \geq 2$ and set

$$q := 2^B.$$

Then $q \geq 4$.

Membership in NP. Given a finite game

$$M = (\mathcal{X}, \mathcal{Y}, \mathcal{A}, p, u),$$

its safe-confusability graph G_M can be computed in polynomial time by scanning all pairs $y, y' \in \mathcal{Y}$ and checking whether there exists $x \in \mathcal{X}$ with

$$p(x, y)p(x, y') > 0 \qquad \text{and} \qquad a^\star(x, y) \neq a^\star(x, y').$$

For fixed $q = 2^B$, a certificate that $\chi(G_M) \leq q$ is simply a map

$$c : \mathcal{Y} \to [q],$$

which can be verified in polynomial time by checking that adjacent vertices of G_M receive distinct colors. By Theorem 4.1,

$$\mathrm{SRC}_{\mathrm{FL}}(M) \leq B \iff \chi(G_M) \leq q.$$

Therefore $\textsc{Src}_{\leq B} \in \mathrm{NP}$.

NP-hardness. We reduce from GRAPH-3-COLORABILITY, which is NP-complete (Garey et al., 1976). Let G be an arbitrary input graph. Form the graph $H := J_q(G)$ as in Theorem E.4. By Theorem E.4,

$$G \text{ is 3-colorable} \iff H \text{ is } q\text{-colorable}.$$

Now construct the graph-realization game M_H from Theorem E.1. This construction is polynomial in the size of G.

By Theorem E.3,

$$\mathrm{SRC}_{\mathrm{FL}}(M_H) \leq B \iff \lceil \log_2 \chi(H) \rceil \leq B.$$

Since $q = 2^B$, the latter is equivalent to

$$\chi(H) \leq 2^B = q,$$

which in turn is equivalent to

$$H \text{ is } q\text{-colorable}.$$

Combining the equivalences, we obtain

$$G \text{ is 3-colorable} \iff \mathrm{SRC}_{\mathrm{FL}}(M_H) \leq B.$$

Thus GRAPH-3-COLORABILITY many-one reduces in polynomial time to $\textsc{Src}_{\leq B}$, proving NP-hardness.

Since the reduction outputs only graph-realization games of the form M_H, the NP-hardness statement already holds on that restricted subclass. □

The complexity of exact computation follows immediately.

Corollary E.6 (Exact computation is NP-hard)**.** *The problem of computing* $\mathrm{SRC}_{\mathrm{FL}}(M)$ *exactly is NP-hard. More strongly, for every fixed integer* $B \geq 2$*, deciding whether*

$$\mathrm{SRC}_{\mathrm{FL}}(M) \leq B$$

is NP-complete.

Proof. The second statement is exactly Theorem E.5. The first follows immediately: if $\mathrm{SRC}_{\mathrm{FL}}(M)$ could be computed exactly in polynomial time, then for any fixed $B \geq 2$ one could decide whether $\mathrm{SRC}_{\mathrm{FL}}(M) \leq B$ in polynomial time, contradicting Theorem E.5 unless $\mathrm{P} = \mathrm{NP}$. □

Remark E.7 (Structural completeness). Theorem E.2 shows that the safe-confusability formalism is not restricted to some narrow subclass of graph-coloring phenomena. Every finite graph appears exactly as a trustworthy-coordination conflict pattern of a one-shot directional assistance game. Consequently, every lower bound, extremal construction, or complexity phenomenon that depends only on proper vertex colorings can be transported into the fixed-length safe-revelation problem.

F. Covert-Channel Collapse Proof

This appendix proves the paper's main trust-separation theorem. The result shows that explicit-message auditing is fundamentally incomplete: even when every available observation-modulation action is individually *less informative* than an approved baseline observation channel in the sense of Blackwell (Blackwell, 1953), the sender can still encode the missing coordination information in the *choice* of which less-informative channel to apply. This formalizes, in a one-shot theorem, the kind of observation-interference concern isolated by Emmons et al. (2025).

Throughout this appendix, fix a one-shot directional assistance game

$$M = (\mathcal{X}, \mathcal{Y}, \mathcal{A}, p, u)$$

satisfying Theorem B.2, and fix a proper coloring

$$c : \mathcal{Y} \to [q]$$

of the safe-confusability graph G_M, for some integer $q \geq 1$.

F.1. Covert Observation-Channel Augmentation

We begin by defining the augmented game. Intuitively, the baseline observation kernel reveals the receiver's original private observation together with an auxiliary nuisance bit. Each covert kernel deletes the nuisance bit and removes one labeled token from a publicly known list of q tokens. The lost nuisance bit makes each covert kernel strictly Blackwell-inferior to the baseline, while the identity of the missing token reveals the sender-chosen color class.

Definition F.1 (Covert observation-channel augmentation)**.** Let

$$\mathcal{R} := \{0, 1\}$$

be a nuisance-bit alphabet, and let

$$\mathcal{Z} := \{0\} \cup [q]$$

be the set of observation-modulation actions. Define the augmented observation alphabet

$$\mathcal{O} := \mathcal{X} \times (\mathcal{R} \cup \{\perp\}) \times 2^{[q]},$$

where $\perp$ is a symbol not in $\mathcal{R}$.

The *covert observation-channel augmentation* of M relative to the coloring c is the one-shot game

$$\widetilde{M}_c = (\mathcal{S}, \mathcal{Y}, \mathcal{Z}, \mathcal{O}, \mathcal{A}, \widetilde{p}, \{Q_z\}_{z \in \mathcal{Z}}, \widetilde{u}),$$

defined as follows:

1. The latent receiver state space is
$$\mathcal{S} := \mathcal{X} \times \mathcal{R}.$$
2. Nature draws
$$(X, Y) \sim p, \qquad R \sim \mathrm{Unif}(\mathcal{R}),$$
independently, and sets the latent receiver state to
$$S := (X, R) \in \mathcal{S}.$$
Equivalently,
$$\widetilde{p}((x, r), y) = \frac{1}{2}\, p(x, y) \qquad \forall (x, r) \in \mathcal{S},\ y \in \mathcal{Y}.$$
3. The sender observes Y and chooses an observation-modulation action $Z \in \mathcal{Z}$.
4. Conditional on the latent receiver state (x, r) and the chosen observation-modulation action $z \in \mathcal{Z}$, the receiver observes an output $O \in \mathcal{O}$ according to the kernel Q_z defined by:
$$Q_0(\cdot \mid x, r) = \delta_{(x, r, [q])},$$
$$Q_m(\cdot \mid x, r) = \delta_{(x, \perp, [q] \setminus \{m\})} \qquad \forall m \in [q].$$

5. After observing O, the receiver chooses an action $a \in \mathcal{A}$.

6. The common payoff is
$$\widetilde{u}(a, (x, r), y, z) := u(a, x, y).$$

Thus the sender's observation-modulation choice affects only the receiver's observation channel and never the payoff directly. In particular, there is no state transition in the one-shot setting, so all observation-modulation actions are payoff-equivalent.

We will need two derived architectures.

Definition F.2 (Approved-baseline safe architecture)**.** The *approved-baseline safe architecture* associated with $\widetilde{M_c}$ is the one-shot directional assistance game
$$M_c^{\text{base}} = (\widetilde{\mathcal{X}}, \mathcal{Y}, \mathcal{A}, p_c^{\text{base}}, u_c^{\text{base}})$$
obtained by forcing the sender to use the baseline observation-modulation action $Z = 0$ and allowing coordination only through an explicit message channel. Here
$$\widetilde{\mathcal{X}} := \mathcal{X} \times \mathcal{R} \times \{[q]\},$$
$$p_c^{\text{base}}((x, r, [q]), y) = \frac{1}{2}\, p(x, y),$$
and
$$u_c^{\text{base}}(a, (x, r, [q]), y) := u(a, x, y).$$

Definition F.3 (Unrestricted covert architecture)**.** The *unrestricted covert architecture* associated with $\widetilde{M_c}$ allows the sender to choose any observation-modulation action $Z \in \mathcal{Z}$ as a function of Y, but does not use any explicit message channel.

F.2. Each Covert Kernel Is Strictly Blackwell-Inferior

We first prove that every covert observation-modulation action is individually strictly less informative than the approved baseline kernel.

Lemma F.4 (Strict Blackwell inferiority of the covert kernels)**.** *For every $m \in [q]$,*
$$Q_m \prec_{\mathrm{B}} Q_0$$
as kernels on the common state space
$$\mathcal{S} = \mathcal{X} \times \mathcal{R}.$$

Proof. Fix $m \in [q]$.

Step 1: $Q_m \preceq_{\mathrm{B}} Q_0$. Define a deterministic garbling
$$K_m : \mathcal{O} \to \Delta(\mathcal{O})$$
by
$$K_m\big((x', b', T') \mid (x, b, T)\big) := \mathbf{1}\{x' = x,\ b' = \perp,\ T' = [q] \setminus \{m\}\}.$$
That is, regardless of the nuisance bit b and regardless of the token set T, the garbling keeps the first coordinate x and outputs the fixed pair
$$(\perp, [q] \setminus \{m\})$$
in the remaining coordinates. For every latent state $(x, r) \in \mathcal{S}$,
$$Q_0(\cdot \mid x, r) = \delta_{(x, r, [q])},$$
so
$$(K_m \circ Q_0)(\cdot \mid x, r) = \delta_{(x, \perp, [q] \setminus \{m\})} = Q_m(\cdot \mid x, r).$$
Hence
$$Q_m \preceq_{\mathrm{B}} Q_0.$$

Step 2: $Q_0 \not\preceq_{\mathrm{B}} Q_m$**.** Consider the Bayesian decision problem whose latent state is

$$S = (X, R) \in \mathcal{X} \times \mathcal{R},$$

whose action set is

$$\mathcal{A}_{\mathrm{dec}} := \mathcal{R},$$

and whose payoff is

$$\phi(\widehat{r}, (x, r)) := \mathbf{1}\{\widehat{r} = r\}.$$

In words, the decision-maker is rewarded for correctly guessing the nuisance bit R.

Under the baseline kernel Q_0, the observation is

$$O_0 = (X, R, [q]),$$

so R is observed perfectly. Therefore the optimal expected payoff equals

$$\sup_{\pi:\mathcal{O}\to\mathcal{R}} \mathbb{E}\big[\phi(\pi(O_0), S)\big] = 1.$$

Under the covert kernel Q_m, the observation is

$$O_m = (X, \perp, [q] \setminus \{m\}),$$

which contains no information about R. Since R is independent of X and is uniform on $\{0, 1\}$, every decision rule $\pi : \mathcal{O} \to \mathcal{R}$ achieves expected payoff at most

$$\max\{\mathbb{P}(R = 0), \mathbb{P}(R = 1)\} = \frac{1}{2}.$$

Thus

$$\sup_{\pi:\mathcal{O}\to\mathcal{R}} \mathbb{E}\big[\phi(\pi(O_m), S)\big] = \frac{1}{2} < 1.$$

By Blackwell's theorem (Blackwell, 1953), this strict separation in decision value implies

$$Q_0 \not\preceq_{\mathrm{B}} Q_m.$$

Combining the two steps yields

$$Q_m \prec_{\mathrm{B}} Q_0.$$

□

The previous lemma is the core trust claim at the observation level: each covert action is individually dominated by the approved baseline as an experiment on the receiver's latent state.

F.3. The Approved-Baseline Architecture Preserves Safe Revelation Complexity

We next show that, if observation modulation is forbidden and the sender is forced to use the approved baseline kernel Q_0, then the explicit safe communication problem is exactly the same as in the original game M.

Lemma F.5 (Baseline graph preservation)**.** *The safe-confusability graph of the approved-baseline architecture M_c^{base} is exactly G_M. Consequently,*

$$\mathrm{SRC}_{\mathrm{FL}}(M_c^{\mathrm{base}}) = \mathrm{SRC}_{\mathrm{FL}}(M) = \lceil \log_2 \chi(G_M) \rceil .$$

Proof. We first identify the optimal action rule in M_c^{base}. A support point of M_c^{base} has the form

$$((x, r, [q]), y)$$

with $p(x, y) > 0$ and $r \in \mathcal{R}$. Since

$$u_c^{\mathrm{base}}(a, (x, r, [q]), y) = u(a, x, y),$$

the maximizing action is unchanged:

$$a^\star_{\text{base}}((x, r, [q]), y) = a^\star(x, y).$$

Now let G_{base} denote the safe-confusability graph of M_c^{base}. By definition, for distinct $y, y' \in \mathcal{Y}$,

$$\{y, y'\} \in E(G_{\text{base}})$$

if and only if there exists some receiver observation

$$\widetilde{x} = (x, r, [q]) \in \widetilde{\mathcal{X}}$$

such that

$$p_c^{\text{base}}(\widetilde{x}, y)\, p_c^{\text{base}}(\widetilde{x}, y') > 0$$

and

$$a^\star_{\text{base}}(\widetilde{x}, y) \neq a^\star_{\text{base}}(\widetilde{x}, y').$$

Using

$$p_c^{\text{base}}((x, r, [q]), y) = \frac{1}{2}\, p(x, y)$$

and

$$a^\star_{\text{base}}((x, r, [q]), y) = a^\star(x, y),$$

this condition is equivalent to the existence of some $x \in \mathcal{X}$ and some $r \in \mathcal{R}$ such that

$$p(x, y)p(x, y') > 0 \qquad \text{and} \qquad a^\star(x, y) \neq a^\star(x, y').$$

Since the nuisance bit r is unrestricted and plays no role in the action comparison, the latter is equivalent to

$$\exists x \in \mathcal{X} \text{ such that } p(x, y)p(x, y') > 0 \text{ and } a^\star(x, y) \neq a^\star(x, y'),$$

which is exactly the edge condition for G_M. Therefore

$$G_{\text{base}} = G_M.$$

Applying Theorem 4.1 to M_c^{base} and to M yields

$$\text{SRC}_{\text{FL}}(M_c^{\text{base}}) = \lceil \log_2 \chi(G_{\text{base}}) \rceil = \lceil \log_2 \chi(G_M) \rceil = \text{SRC}_{\text{FL}}(M).$$

□

Thus the covert augmentation does not make the safe explicit-message problem any easier when the observation channel is fixed to the approved baseline.

F.4. Zero Explicit Bits Suffice in the Unrestricted Architecture

We now show that, once the sender may choose among the covert kernels, the missing color class can be communicated with zero explicit message bits.

For each $x \in \mathcal{X}$ and each color $m \in [q]$, define

$$\mathcal{Y}(x, m) := \{y \in \mathcal{Y} : p(x, y) > 0,\ c(y) = m\}.$$

Lemma F.6 (Color-conditioned optimal action is well-defined)**.** *Fix $(x, m) \in \mathcal{X} \times [q]$. If $\mathcal{Y}(x, m) \neq \varnothing$, then the value $a^\star(x, y)$ is the same for all $y \in \mathcal{Y}(x, m)$.*

Proof. Let $y, y' \in \mathcal{Y}(x, m)$. Then

$$p(x, y) > 0, \qquad p(x, y') > 0, \qquad c(y) = c(y') = m.$$

If $a^\star(x, y) \neq a^\star(x, y')$, then by definition of the safe-confusability graph,

$$\{y, y'\} \in E_M.$$

But c is a proper coloring of G_M, so adjacent vertices cannot share the same color. This contradicts $c(y) = c(y')$. Hence

$$a^\star(x, y) = a^\star(x, y') \qquad \forall y, y' \in \mathcal{Y}(x, m).$$

□

We can now construct the covert decoder.

Lemma F.7 (Zero-explicit-message covert decoder)**.** *There exists a deterministic decoder*

$$d^{\text{cov}} : \mathcal{O} \to \mathcal{A}$$

such that, if the sender chooses the observation-modulation action

$$Z = c(Y)$$

and sends no explicit message, then

$$d^{\text{cov}}(O) = a^\star(X, Y) \qquad \textit{almost surely}.$$

Consequently, the unrestricted covert architecture achieves value $V^\star(M)$ with zero explicit message bits.

Proof. For every observation

$$o = (x, b, T) \in \mathcal{O},$$

define $d^{\text{cov}}(o)$ as follows.

If

$$T = [q] \setminus \{m\}$$

for some unique $m \in [q]$ and $\mathcal{Y}(x, m) \neq \varnothing$, define

$$d^{\text{cov}}(x, b, T) := a^\star(x, y)$$

for any $y \in \mathcal{Y}(x, m)$. This is well-defined by Theorem F.6. In all other cases, define $d^{\text{cov}}(x, b, T)$ arbitrarily.

Now consider the on-path evolution under the zero-explicit-message policy

$$Z = c(Y).$$

Fix any support point with

$$p(x, y) > 0 \qquad \text{and} \qquad r \in \mathcal{R}.$$

The chosen observation-modulation action is

$$z = c(y) \in [q].$$

By definition of the covert kernel,

$$O = (x, \perp, [q] \setminus \{c(y)\}) \qquad \text{almost surely.}$$

Since $p(x, y) > 0$ and $c(y) = z$, we have

$$y \in \mathcal{Y}(x, z),$$

so $\mathcal{Y}(x, z) \neq \varnothing$. Therefore the decoder definition gives

$$d^{\text{cov}}(O) = a^\star(x, y).$$

Because this holds for every support point (x, y) and every nuisance bit r, we obtain

$$d^{\text{cov}}(O) = a^\star(X, Y) \qquad \text{almost surely.}$$

Hence the achieved expected payoff is

$$\mathbb{E}\big[u(d^{\text{cov}}(O), X, Y)\big] = \mathbb{E}\big[u(a^\star(X, Y), X, Y)\big] = V^\star(M).$$

No explicit message channel is used anywhere in this construction, so the protocol uses zero explicit message bits. □

F.5. Main Collapse Theorem

We can now combine the preceding ingredients into the promised theorem.

Theorem F.8 (Covert-channel collapse)**.** *Let*

$$M = (\mathcal{X}, \mathcal{Y}, \mathcal{A}, p, u)$$

satisfy Theorem B.2, and let

$$c : \mathcal{Y} \to [q]$$

be any proper coloring of the safe-confusability graph G_M. Then there exists a covert observation-channel augmentation $\widetilde{M}_c$ of M such that:

1. *for every covert observation-modulation action $m \in [q]$,*

$$Q_m \prec_{\mathrm{B}} Q_0;$$

2. *in the approved-baseline safe architecture,*

$$\mathrm{SRC}_{\mathrm{FL}}(M_c^{\mathrm{base}}) = \mathrm{SRC}_{\mathrm{FL}}(M) = \lceil \log_2 \chi(G_M) \rceil ;$$

3. *in the unrestricted covert architecture, zero explicit message bits suffice to attain the full-information value: there exists a zero-explicit-message policy with achieved value*

$$V^\star(M).$$

Proof. Part (1) is exactly Theorem F.4. Part (2) is exactly Theorem F.5. Part (3) is exactly Theorem F.7. □

Corollary F.9 (Collapse at the chromatic threshold)**.** *If c is chosen to be an optimal coloring of G_M, so that $q = \chi(G_M)$, then the same augmentation exhibits a collapse from the exact safe explicit-message requirement*

$$\lceil \log_2 \chi(G_M) \rceil$$

to zero explicit bits in the unrestricted covert architecture.

Proof. Immediate from Theorem F.8. □

Remark F.10 (Why the theorem is nontrivial). The theorem does *not* say that a more informative observation channel can replace explicit communication. Quite the opposite: every covert kernel Q_m is individually strictly *worse* than the approved baseline Q_0 as an experiment on the receiver's latent state. The separation comes entirely from the sender's freedom to choose *which* inferior kernel to apply. That choice acts as a policy-level communication channel whose capacity is invisible to audits that inspect only the informativeness of each observation action in isolation.

G. Missing-Bit Family Proofs

This appendix introduces an explicit family of one-shot directional assistance games for which the entire fixed-budget value curve can be computed in closed form. The family plays two roles. First, it gives a concrete extremal witness for the fixed-length theorem: the safe-confusability graph is a complete graph on 2^k vertices, so exact safe optimality requires exactly k bits. Second, and more importantly for the present paper, it shows that losing even a single safe bit can incur a constant drop in the maximum attainable safe value. In fact, on this family, every lost bit halves the best achievable safe payoff.

G.1. Definition of the Family

Definition G.1 (Missing-bit family)**.** For every integer $k \geq 1$, define the one-shot directional assistance game

$$M_k = (\mathcal{X}_k, \mathcal{Y}_k, \mathcal{A}_k, p_k, u_k)$$

by

$$\mathcal{X}_k := \{x_0\}, \qquad \mathcal{Y}_k := \mathcal{A}_k := \{1, \ldots, 2^k\},$$
$$p_k(x_0, y) := 2^{-k} \qquad \forall y \in \mathcal{Y}_k,$$

and

$$u_k(a, x_0, y) := \mathbf{1}\{a = y\} \qquad \forall a \in \mathcal{A}_k,\ y \in \mathcal{Y}_k.$$

Thus the receiver has no private side information beyond a constant observation, the sender privately observes a uniformly random symbol $Y \in \{1, \ldots, 2^k\}$, and the team's payoff is one if and only if the receiver outputs the sender's symbol exactly.

Lemma G.2 (Basic structure of M_k)**.** *For every $k \geq 1$, the game M_k satisfies Theorem B.2, and for every $y \in \mathcal{Y}_k$,*

$$a^\star(x_0, y) = y.$$

Moreover, the safe-confusability graph of M_k is the complete graph

$$G_{M_k} = K_{2^k}.$$

Proof. Fix $k \geq 1$. For every $y \in \mathcal{Y}_k$ and every action $a \in \mathcal{A}_k$,

$$u_k(a, x_0, y) = \mathbf{1}\{a = y\}.$$

Hence

$$u_k(y, x_0, y) = 1 \qquad \text{and} \qquad u_k(a, x_0, y) = 0 \ \text{ for all } a \neq y.$$

Therefore the maximizer of $a \mapsto u_k(a, x_0, y)$ is unique and equal to y. This proves the strict-optimality condition and the identity $a^\star(x_0, y) = y$.

Now let $y \neq y'$ be any two distinct elements of $\mathcal{Y}_k$. Since $p_k(x_0, y) = p_k(x_0, y') = 2^{-k} > 0$ and

$$a^\star(x_0, y) = y \neq y' = a^\star(x_0, y'),$$

the definition of the safe-confusability graph gives

$$\{y, y'\} \in E_{M_k}.$$

Since this holds for every distinct pair y, y', the graph G_{M_k} is the complete graph on 2^k vertices. □

Corollary G.3 (Exact safe-optimality threshold)**.** *For every $k \geq 1$,*

$$\mathrm{SRC}_{\mathrm{FL}}(M_k) = k.$$

Equivalently, the minimum number of explicit safe messages required to attain the full-information value in M_k is exactly 2^k.

Proof. By Theorem G.2, $G_{M_k} = K_{2^k}$, whose chromatic number is

$$\chi(G_{M_k}) = 2^k.$$

Applying Theorem 4.1 yields

$$\mathrm{SRC}_{\mathrm{FL}}(M_k) = \lceil \log_2 \chi(G_{M_k}) \rceil = \lceil \log_2 2^k \rceil = k.$$

The equivalent message-alphabet statement is the second part of Theorem 4.1. □

G.2. Exact Fixed-Budget Value Curve

The main novelty of this family is not merely that the exact safe-optimality threshold equals k, but that the best value achievable below that threshold can also be characterized exactly.

Fix $k \geq 1$ and consider any deterministic safe protocol

$$e : \mathcal{Y}_k \to \mathcal{M}, \qquad d : \mathcal{X}_k \times \mathcal{M} \to \mathcal{A}_k.$$

Since $\mathcal{X}_k = \{x_0\}$ is a singleton, the decoder is equivalently a map

$$g : \mathcal{M} \to \mathcal{A}_k, \qquad g(m) := d(x_0, m).$$

For each message $m \in \mathcal{M}$, define the corresponding message class

$$C_m := e^{-1}(m) \subseteq \mathcal{Y}_k.$$

Lemma G.4 (Partition formula)**.** *For every deterministic safe protocol (e, d) on M_k,*

$$V_{M_k}(e, d) = 2^{-k} \sum_{m \in \mathcal{M}} \sum_{y \in C_m} \mathbf{1}\{g(m) = y\}.$$

Equivalently,

$$V_{M_k}(e, d) = 2^{-k} \sum_{m \in \mathcal{M}} \mathbf{1}\{g(m) \in C_m\}.$$

In particular,

$$V_{M_k}(e, d) \leq 2^{-k} \left|\{m \in \mathcal{M} : C_m \neq \varnothing\}\right|.$$

Proof. By definition of M_k,

$$V_{M_k}(e, d) = \mathbb{E}_{p_k}[u_k(d(X, e(Y)), X, Y)] = 2^{-k} \sum_{y \in \mathcal{Y}_k} \mathbf{1}\{d(x_0, e(y)) = y\}.$$

Since $g(m) = d(x_0, m)$ and the message classes partition $\mathcal{Y}_k$, we may rewrite this as

$$V_{M_k}(e, d) = 2^{-k} \sum_{m \in \mathcal{M}} \sum_{y \in C_m} \mathbf{1}\{g(m) = y\}.$$

For a fixed message m, the inner sum is at most one because $g(m)$ is a single action in $\mathcal{A}_k$. Moreover, the inner sum equals $\mathbf{1}\{g(m) \in C_m\}$. This proves both displayed identities and the upper bound. □

We can now solve the budget-constrained optimization problem exactly.

Theorem G.5 (Exact budget-value tradeoff for the missing-bit family)**.** *Fix $k \geq 1$. Then for every integer $B \geq 0$,*

$$\sup \left\{V_{M_k}(e, d) : (e, d) \text{ is a deterministic safe protocol with } |\mathcal{M}| \leq 2^B\right\} = \min\{1, 2^{B-k}\}.$$

Equivalently, the safe-value gap satisfies

$$\Delta_{M_k}(B) = 1 - \min\{1, 2^{B-k}\}.$$

In particular, for every integer $0 \leq B \leq k$,

$$\sup_{|\mathcal{M}| \leq 2^B} V_{M_k}(e, d) = 2^{B-k} \qquad \text{and} \qquad \Delta_{M_k}(B) = 1 - 2^{B-k}.$$

Proof. Fix $k \geq 1$ and an integer $B \geq 0$. Set

$$q := 2^B.$$

Upper bound. Let (e, d) be any deterministic safe protocol with $|\mathcal{M}| \leq q$. By Theorem G.4,

$$V_{M_k}(e, d) \leq 2^{-k} \left|\{m \in \mathcal{M} : C_m \neq \varnothing\}\right| \leq 2^{-k} \min\{|\mathcal{M}|, 2^k\} \leq 2^{-k} \min\{q, 2^k\}.$$

Since $q = 2^B$, the final quantity is exactly

$$\min\{1, 2^{B-k}\}.$$

Hence

$$\sup_{|\mathcal{M}| \leq 2^B} V_{M_k}(e, d) \leq \min\{1, 2^{B-k}\}.$$

Lower bound when $B \geq k$. If $B \geq k$, then $q = 2^B \geq 2^k = |\mathcal{Y}_k|$. Choose a message alphabet of size 2^k, define the encoder by

$$e(y) := y,$$

and define the decoder by

$$g(m) := m.$$

Then

$$g(e(y)) = y \qquad \forall y \in \mathcal{Y}_k,$$

so

$$V_{M_k}(e, d) = 1.$$

Thus

$$\sup_{|\mathcal{M}| \leq 2^B} V_{M_k}(e, d) \geq 1 = \min\{1, 2^{B-k}\}.$$

Lower bound when $B < k$. Assume now $B < k$, so $q = 2^B < 2^k$. Let the message alphabet be

$$\mathcal{M} = \{1, \ldots, q\}.$$

Define the encoder by

$$e(y) := \begin{cases} y, & \text{if } 1 \leq y \leq q-1, \\ q, & \text{if } q \leq y \leq 2^k. \end{cases}$$

Thus the message classes are

$$C_m = \{m\} \quad \text{for } 1 \leq m \leq q-1, \qquad C_q = \{q, q+1, \ldots, 2^k\}.$$

Define the decoder by

$$g(m) := m \qquad \forall m \in \{1, \ldots, q\}.$$

Then

$$g(m) \in C_m \qquad \forall m \in \{1, \ldots, q\},$$

so by Theorem G.4,

$$V_{M_k}(e, d) = 2^{-k} \sum_{m=1}^{q} \mathbf{1}\{g(m) \in C_m\} = 2^{-k} q = 2^{B-k}.$$

Hence

$$\sup_{|\mathcal{M}| \leq 2^B} V_{M_k}(e, d) \geq 2^{B-k} = \min\{1, 2^{B-k}\}.$$

Combining the upper and lower bounds proves

$$\sup_{|\mathcal{M}| \leq 2^B} V_{M_k}(e, d) = \min\{1, 2^{B-k}\}.$$

The formula for the safe-value gap follows immediately from Theorem B.5 and the fact that $V^\star(M_k) = 1$. □

Corollary G.6 (The cost of a missing safe bit)**.** *For every $k \geq 1$,*

$$\Delta_{M_k}(k-1) = \frac{1}{2}.$$

More generally, for every integer $\ell \in \{0, 1, \ldots, k\}$,

$$\sup_{|\mathcal{M}| \leq 2^{k-\ell}} V_{M_k}(e, d) = 2^{-\ell} \qquad \textit{and} \qquad \Delta_{M_k}(k-\ell) = 1 - 2^{-\ell}.$$

Thus, on the family $\{M_k\}_{k \geq 1}$, each missing safe bit halves the maximum attainable safe value.

Proof. Set $B = k - 1$ in Theorem G.5 to obtain

$$\sup_{|\mathcal{M}| \leq 2^{k-1}} V_{M_k}(e, d) = 2^{-1} = \frac{1}{2}.$$

Since $V^\star(M_k) = 1$, this implies

$$\Delta_{M_k}(k-1) = 1 - \frac{1}{2} = \frac{1}{2}.$$

The more general statement follows by setting $B = k - \ell$ in the same theorem. □

Remark G.7 (Why this family matters). The family $\{M_k\}$ is intentionally extreme: the receiver has no private information, and every pair of sender symbols is conflicting. Precisely because of this extremality, the family isolates the sharpest possible dependence of safe value on the explicit-message budget. It will therefore serve as the cleanest quantitative benchmark for the main-body trust-separation theorem: once covert observation-side channels are added, the same family will admit full value with zero explicit communication, in stark contrast to the exact budget-value law above.

H. Approximate-Value and Robustness Lemmas

This appendix records several quantitative consequences of the strict-optimality assumption. The results serve two purposes. First, they convert approximate action recovery into approximate value guarantees, both in the one-shot and block-coding regimes. Second, they show that the graph-theoretic objects studied in the main text are stable under sufficiently small uniform perturbations of the payoff function. These lemmas justify the use of approximate numerical solvers in the experiments and clarify the operational meaning of the exact characterizations proved earlier.

H.1. One-Shot Suboptimality Gap Decomposition

Fix a one-shot directional assistance game

$$M = (\mathcal{X}, \mathcal{Y}, \mathcal{A}, p, u)$$

satisfying Theorem B.2. Recall that

$$u^\star(x, y) := \max_{a \in \mathcal{A}} u(a, x, y), \qquad a^\star(x, y) \in \arg\max_{a \in \mathcal{A}} u(a, x, y)$$

denote the full-information optimal payoff and the unique optimal action on $\mathrm{supp}(p)$.

Definition H.1 (Pointwise suboptimality gap)**.** For $(x, y, a) \in \mathcal{X} \times \mathcal{Y} \times \mathcal{A}$, define

$$\mathrm{gap}_M(x, y, a) := u^\star(x, y) - u(a, x, y).$$

Define the minimum support margin and the maximum one-step loss by

$$\underline{\gamma}(M) := \min_{(x,y) \in \mathrm{supp}(p)} \min_{a \in \mathcal{A}: a \neq a^\star(x,y)} \mathrm{gap}_M(x, y, a),$$

and

$$\overline{\gamma}(M) := \max_{(x,y) \in \mathrm{supp}(p)} \max_{a \in \mathcal{A}} \mathrm{gap}_M(x, y, a).$$

Because the alphabets are finite and Theorem B.2 holds, $\underline{\gamma}(M)$ is strictly positive whenever $|\mathcal{A}| \geq 2$. If $|\mathcal{A}| = 1$, then all statements below are trivial, since every protocol is automatically optimal.

Lemma H.2 (Value-gap identity)**.** *Let $\widehat{A}$ be any $\mathcal{A}$-valued random variable defined on the same probability space as $(X, Y) \sim p$. Then*

$$V^\star(M) - \mathbb{E}[u(\widehat{A}, X, Y)] = \mathbb{E}\big[\mathrm{gap}_M(X, Y, \widehat{A})\big].$$

Moreover, on $\mathrm{supp}(p)$*,*

$$\mathrm{gap}_M(x, y, a) = 0 \iff a = a^\star(x, y).$$

Proof. By definition of gap_M,

$$\mathrm{gap}_M(X, Y, \widehat{A}) = u^\star(X, Y) - u(\widehat{A}, X, Y).$$

Taking expectations gives

$$\mathbb{E}\big[\mathrm{gap}_M(X, Y, \widehat{A})\big] = \mathbb{E}[u^\star(X, Y)] - \mathbb{E}[u(\widehat{A}, X, Y)] = V^\star(M) - \mathbb{E}[u(\widehat{A}, X, Y)].$$

This proves the identity.

For the second claim, nonnegativity of gap_M is immediate from the definition of $u^\star$. If $(x, y) \in \mathrm{supp}(p)$ and $\mathrm{gap}_M(x, y, a) = 0$, then $u(a, x, y) = u^\star(x, y)$, so a is a maximizer of $a' \mapsto u(a', x, y)$. By Theorem B.2, the maximizer is unique, hence $a = a^\star(x, y)$. The converse is immediate. □

The next corollary converts action error probability into value loss and vice versa.

Corollary H.3 (Action-error/value-gap sandwich)**.** *Let $\widehat{A}$ be any $\mathcal{A}$-valued random variable defined on the same probability space as $(X, Y) \sim p$, and define the action-error event*

$$E := \{\widehat{A} \neq a^\star(X, Y)\}, \qquad p_e := \mathbb{P}(E).$$

Then

$$\underline{\gamma}(M)\, p_e \;\leq\; V^\star(M) - \mathbb{E}[u(\widehat{A}, X, Y)] \;\leq\; \overline{\gamma}(M)\, p_e.$$

Equivalently,

$$p_e \;\leq\; \frac{V^\star(M) - \mathbb{E}[u(\widehat{A}, X, Y)]}{\underline{\gamma}(M)}$$

and

$$V^\star(M) - \mathbb{E}[u(\widehat{A}, X, Y)] \;\leq\; \overline{\gamma}(M)\, p_e.$$

Proof. By Theorem H.2,

$$V^\star(M) - \mathbb{E}[u(\widehat{A}, X, Y)] = \mathbb{E}\big[\mathrm{gap}_M(X, Y, \widehat{A})\big].$$

On the event E^c, Theorem H.2 gives $\mathrm{gap}_M(X, Y, \widehat{A}) = 0$. On the event E, the action $\widehat{A}$ is suboptimal, so by definition of $\underline{\gamma}(M)$ and $\overline{\gamma}(M)$,

$$\underline{\gamma}(M) \leq \mathrm{gap}_M(X, Y, \widehat{A}) \leq \overline{\gamma}(M) \qquad \text{almost surely on } E.$$

Therefore

$$\underline{\gamma}(M)\mathbf{1}_E \leq \mathrm{gap}_M(X, Y, \widehat{A}) \leq \overline{\gamma}(M)\mathbf{1}_E \qquad \text{almost surely.}$$

Taking expectations yields the stated inequality. □

Remark H.4 (Exact coincidence for the missing-bit family). For the family M_k of Theorem G.1, one has

$$\underline{\gamma}(M_k) = \overline{\gamma}(M_k) = 1.$$

Hence

$$V^\star(M_k) - \mathbb{E}[u(\widehat{A}, X, Y)] = \mathbb{P}\{\widehat{A} \neq Y\}.$$

Thus, on M_k, the value gap is exactly the action-error probability.

H.2. Block Codes and Average Payoff

We now pass to the repeated game with average payoff, using the notation of Section D. Let

$$\widehat{A}^n = (\widehat{A}_1, \dots, \widehat{A}_n) \in \mathcal{A}^n$$

be any random block action, possibly produced by a block decoder that depends on the entire pair (X^n, Y^n) only through X^n and an encoded message.

Proposition H.5 (Block value-gap identity)**.** *For every $n \geq 1$,*

$$V^\star(M^{\otimes n}) - \mathbb{E}\big[u_n(\widehat{A}^n, X^n, Y^n)\big] = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\big[\mathrm{gap}_M(X_i, Y_i, \widehat{A}_i)\big].$$

Equivalently, if

$$E_i := \{\widehat{A}_i \neq a^\star(X_i, Y_i)\}, \qquad \bar{p}_e^{(n)} := \frac{1}{n}\sum_{i=1}^{n} \mathbb{P}(E_i),$$

then

$$\underline{\gamma}(M)\,\bar{p}_e^{(n)} \;\leq\; V^\star(M^{\otimes n}) - \mathbb{E}\big[u_n(\widehat{A}^n, X^n, Y^n)\big] \;\leq\; \overline{\gamma}(M)\,\bar{p}_e^{(n)}.$$

Proof. By definition of the repeated-game payoff,

$$u_n(\widehat{A}^n, X^n, Y^n) = \frac{1}{n}\sum_{i=1}^{n} u(\widehat{A}_i, X_i, Y_i).$$

Also,

$$V^\star(M^{\otimes n}) = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} u^\star(X_i, Y_i)\right] = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}[u^\star(X_i, Y_i)].$$

Subtracting yields

$$V^\star(M^{\otimes n}) - \mathbb{E}[u_n(\widehat{A}^n, X^n, Y^n)] = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\big[u^\star(X_i, Y_i) - u(\widehat{A}_i, X_i, Y_i)\big].$$

By definition of gap_M, this is exactly

$$\frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\big[\mathrm{gap}_M(X_i, Y_i, \widehat{A}_i)\big].$$

This proves the first identity.

Applying Theorem H.3 separately to each coordinate i gives

$$\underline{\gamma}(M)\,\mathbb{P}(E_i) \leq \mathbb{E}\big[\mathrm{gap}_M(X_i, Y_i, \widehat{A}_i)\big] \leq \overline{\gamma}(M)\,\mathbb{P}(E_i).$$

Averaging over $i \in [n]$ proves the stated bounds. □

The asymptotic coding theorem in the main text is formulated in terms of *block error probability*. The next corollary translates that quantity into value loss.

Corollary H.6 (From block error to value loss)**.** *Let*

$$P_b^{(n)} := \mathbb{P}\Big\{\widehat{A}^n \neq f_M^{\otimes n}(X^n, Y^n)\Big\} = \mathbb{P}\left\{\bigcup_{i=1}^{n} E_i\right\}.$$

Then

$$V^\star(M^{\otimes n}) - \mathbb{E}\big[u_n(\widehat{A}^n, X^n, Y^n)\big] \leq \overline{\gamma}(M)\,P_b^{(n)}.$$

In particular, if $P_b^{(n)} \to 0$, then the achieved average value converges to the full-information optimum:

$$\mathbb{E}\big[u_n(\widehat{A}^n, X^n, Y^n)\big] \to V^\star(M^{\otimes n}) = V^\star(M).$$

Proof. Since

$$\frac{1}{n}\sum_{i=1}^{n}\mathbf{1}_{E_i} \le \mathbf{1}_{\cup_{i=1}^{n} E_i} \qquad \text{almost surely,}$$

we have

$$\bar{p}_e^{(n)} = \frac{1}{n}\sum_{i=1}^{n}\mathbb{P}(E_i) \le \mathbb{P}\left\{\bigcup_{i=1}^{n} E_i\right\} = P_b^{(n)}.$$

Combining this with Theorem H.5 yields

$$V^\star(M^{\otimes n}) - \mathbb{E}[u_n(\widehat{A}^n, X^n, Y^n)] \le \overline{\gamma}(M)\,\bar{p}_e^{(n)} \le \overline{\gamma}(M)\,P_b^{(n)}.$$

If $P_b^{(n)} \to 0$, the right-hand side converges to 0, proving the limit. The identity $V^\star(M^{\otimes n}) = V^\star(M)$ follows from the fact that the block payoff is the coordinatewise average and (X_i, Y_i) are i.i.d. □

H.3. Uniform Payoff Perturbations

We now show that the exact graph-theoretic objects of the paper are stable under sufficiently small uniform perturbations of the payoff function.

Consider two one-shot directional assistance games on the same alphabets and the same observation law,

$$M = (\mathcal{X}, \mathcal{Y}, \mathcal{A}, p, u), \qquad \widetilde{M} = (\mathcal{X}, \mathcal{Y}, \mathcal{A}, p, \widetilde{u}),$$

and write

$$\|u - \widetilde{u}\|_\infty := \max_{(a,x,y)\in\mathcal{A}\times\mathcal{X}\times\mathcal{Y}} |u(a,x,y) - \widetilde{u}(a,x,y)|.$$

Proposition H.7 (Payoff perturbation stability)**.** *Assume M satisfies Theorem B.2 and*

$$\|u - \widetilde{u}\|_\infty < \frac{\underline{\gamma}(M)}{2}.$$

Then:

1. *$\widetilde{M}$ also satisfies Theorem B.2;*
2. *the support-wise optimal action rule is unchanged:*
$$\widetilde{a}^\star(x,y) = a^\star(x,y) \qquad \forall (x,y) \in \mathrm{supp}(p);$$
3. *the safe-confusability graph is unchanged:*
$$G_{\widetilde{M}} = G_M;$$
4. *the fixed-length and asymptotic safe revelation complexities are unchanged:*
$$\mathrm{SRC}_{\mathrm{FL}}(\widetilde{M}) = \mathrm{SRC}_{\mathrm{FL}}(M), \qquad \mathrm{SRC}_\infty(\widetilde{M}) = \mathrm{SRC}_\infty(M);$$
5. *the full-information values differ by at most the perturbation size:*
$$|V^\star(\widetilde{M}) - V^\star(M)| \le \|u - \widetilde{u}\|_\infty.$$

Proof. Fix any support point $(x,y) \in \mathrm{supp}(p)$ and any action $a \ne a^\star(x,y)$. By definition of $\underline{\gamma}(M)$,

$$u(a^\star(x,y), x, y) - u(a, x, y) \ge \underline{\gamma}(M).$$

Therefore

$$\begin{aligned}\widetilde{u}(a^\star(x,y), x, y) - \widetilde{u}(a, x, y) &\ge u(a^\star(x,y), x, y) - u(a, x, y) - 2\|u - \widetilde{u}\|_\infty \\ &> \underline{\gamma}(M) - 2\cdot\frac{\underline{\gamma}(M)}{2} = 0.\end{aligned}$$

Thus $a^\star(x, y)$ remains strictly better than every competing action under $\widetilde{u}$, proving both strict optimality of $\widetilde{M}$ and the identity

$$\widetilde{a}^\star(x, y) = a^\star(x, y) \qquad \forall (x, y) \in \mathrm{supp}(p).$$

This proves (1) and (2).

For (3), recall that the safe-confusability graph is defined by the condition

$$\{y, y'\} \in E_M \iff \exists x \in \mathcal{X} \text{ such that } p(x, y)p(x, y') > 0 \text{ and } a^\star(x, y) \neq a^\star(x, y').$$

Since p is unchanged and the optimal action rule is unchanged on the support, the edge relation is identical for M and $\widetilde{M}$. Hence

$$G_{\widetilde{M}} = G_M.$$

For (4), apply Theorems 4.1 and 4.2 to M and $\widetilde{M}$, using the graph identity from part (3):

$$\mathrm{SRC}_{\mathrm{FL}}(\widetilde{M}) = \lceil \log_2 \chi(G_{\widetilde{M}}) \rceil = \lceil \log_2 \chi(G_M) \rceil = \mathrm{SRC}_{\mathrm{FL}}(M),$$

and similarly

$$\mathrm{SRC}_\infty(\widetilde{M}) = H_{G_{\widetilde{M}}}(Y \mid X) = H_{G_M}(Y \mid X) = \mathrm{SRC}_\infty(M).$$

For (5), since

$$\big|u(a, x, y) - \widetilde{u}(a, x, y)\big| \leq \|u - \widetilde{u}\|_\infty \qquad \forall (a, x, y),$$

and the optimal actions agree on the support, we have

$$\big|\widetilde{u}(\widetilde{a}^\star(x, y), x, y) - u(a^\star(x, y), x, y)\big| = \big|\widetilde{u}(a^\star(x, y), x, y) - u(a^\star(x, y), x, y)\big| \leq \|u - \widetilde{u}\|_\infty$$

for every support point (x, y). Taking expectations yields

$$|V^\star(\widetilde{M}) - V^\star(M)| \leq \|u - \widetilde{u}\|_\infty.$$

This completes the proof. □

Remark H.8 (Use in experiments). Theorems H.3, H.6 and H.7 justify the three kinds of approximations used in the experimental section: finite-block decoders with small residual block error, approximate numerical solvers whose output can be certified by action error, and floating-point perturbations of the payoff table that are much smaller than the support margin.

I. Sequential Local Lower Bound for Finite-Horizon POAGs

This appendix explains how the one-shot safe-confusability formalism arises as a local necessary condition inside finite-horizon partially observable assistance games. The result is deliberately modest: we do *not* claim a full dynamic characterization of sequential safe revelation complexity. Instead, we prove that whenever a deterministic safe policy attains the full-information sequential benchmark exactly, then at every reachable decision stage its local message partition must realize a proper coloring of an induced one-shot safe-confusability graph. Thus the one-shot chromatic lower bound appears as a pointwise obstruction inside sequential trustworthy coordination.

The formalism below is an abstract sender–receiver slice of a finite-horizon common-payoff partially observable game, in the spirit of the POAG model of Emmons et al. (2025). The abstraction is chosen so that the receiver is the unique payoff-relevant decision-maker at each stage, while the sender can coordinate only through an explicit message.

I.1. Finite-Horizon Directional POAGs

Fix a horizon $H \geq 1$. For each stage $t \in [H]$, let $\mathcal{H}_t$ be a finite set of *public histories* available to both agents at the start of stage t. Let $h_1^\circ \in \mathcal{H}_1$ denote the initial public history. For each stage t and public history $h \in \mathcal{H}_t$, let:

- $\mathcal{X}_t(h)$ be the receiver's private observation alphabet;

- $\mathcal{Y}_t(h)$ be the sender's private observation alphabet;
- $\mathcal{A}_t(h)$ be the receiver action set;
- $p_t^h \in \Delta(\mathcal{X}_t(h) \times \mathcal{Y}_t(h))$ be the joint law of the stage-t private observations conditional on $H_t = h$;
- $K_t^h(\cdot, \cdot \mid x, y, a)$ be a stochastic kernel on $\mathbb{R} \times \mathcal{H}_{t+1}$ specifying the joint law of the stage-t reward R_t and next public history H_{t+1} conditional on $(H_t, X_t, Y_t, A_t) = (h, x, y, a)$.

We assume all alphabets are finite and all rewards are bounded.

At stage t, conditional on $H_t = h$, Nature draws

$$(X_t, Y_t) \sim p_t^h.$$

The sender observes Y_t and emits an explicit message

$$M_t \in \mathcal{M}_t(h),$$

where $\mathcal{M}_t(h)$ is a finite message alphabet chosen as part of the policy. The receiver observes X_t and M_t, chooses an action

$$A_t \in \mathcal{A}_t(h),$$

and then (R_t, H_{t+1}) is drawn from $K_t^h(\cdot, \cdot \mid X_t, Y_t, A_t)$.

Definition I.1 (Deterministic safe policy)**.** A *deterministic safe policy* π consists, for each stage $t \in [H]$ and public history $h \in \mathcal{H}_t$, of:

1. a finite message alphabet $\mathcal{M}_t^\pi(h)$;
2. an encoder
$$e_t^\pi(h, \cdot) : \mathcal{Y}_t(h) \to \mathcal{M}_t^\pi(h);$$
3. a decoder
$$d_t^\pi(h, \cdot, \cdot) : \mathcal{X}_t(h) \times \mathcal{M}_t^\pi(h) \to \mathcal{A}_t(h).$$

The induced stage-t receiver action is

$$A_t^\pi = d_t^\pi\big(h, X_t, e_t^\pi(h, Y_t)\big) \qquad \text{when } H_t = h.$$

I.2. Full-Information Benchmark and Local Continuation Games

We now define the full-information dynamic benchmark. In this benchmark, the receiver at stage t is allowed to condition its action on the full pair (X_t, Y_t) in addition to the public history H_t.

Set

$$V_{H+1}^{\mathrm{FI}}(h) := 0 \qquad \forall h \in \mathcal{H}_{H+1}.$$

For $t = H, H-1, \ldots, 1$, define the full-information Q-function

$$Q_t^{\mathrm{FI}}(h, x, y, a) := \mathbb{E}\big[R_t + V_{t+1}^{\mathrm{FI}}(H_{t+1}) \,\big|\, H_t = h,\ X_t = x,\ Y_t = y,\ A_t = a\big],$$

and the full-information value recursion

$$V_t^{\mathrm{FI}}(h) := \sum_{(x,y) \in \mathcal{X}_t(h) \times \mathcal{Y}_t(h)} p_t^h(x, y) \max_{a \in \mathcal{A}_t(h)} Q_t^{\mathrm{FI}}(h, x, y, a). \tag{7}$$

For a deterministic safe policy π, define analogously

$$V_{H+1}^\pi(h) := 0 \qquad \forall h \in \mathcal{H}_{H+1},$$

and, for $t = H, H-1, \ldots, 1$,

$$V_t^\pi(h) := \sum_{(x,y)} p_t^h(x, y)\, \mathbb{E}\big[R_t + V_{t+1}^\pi(H_{t+1}) \,\big|\, H_t = h,\ X_t = x,\ Y_t = y,\ A_t = A_t^\pi\big]. \tag{8}$$

Definition I.2 (Reachable public history)**.** A public history $h \in \mathcal{H}_t$ is *reachable under* π if

$$\mathbb{P}_\pi\{H_t = h\} > 0$$

when the sequential game is started from $H_1 = h_1^\circ$ and both agents follow π.

Fix a stage t and a public history $h \in \mathcal{H}_t$. We associate with (t, h) a one-shot directional assistance game by folding the full-information continuation value into the stage payoff.

Definition I.3 (Local continuation game)**.** The *local continuation game* at stage t and public history h is the one-shot directional assistance game

$$M_{t,h}^{\mathrm{loc}} = \big(\mathcal{X}_t(h),\, \mathcal{Y}_t(h),\, \mathcal{A}_t(h),\, p_t^h,\, u_{t,h}^{\mathrm{loc}}\big),$$

where

$$u_{t,h}^{\mathrm{loc}}(a, x, y) := Q_t^{\mathrm{FI}}(h, x, y, a).$$

Let

$$a_{t,h}^\star(x, y) \in \arg\max_{a \in \mathcal{A}_t(h)} u_{t,h}^{\mathrm{loc}}(a, x, y) = \arg\max_{a \in \mathcal{A}_t(h)} Q_t^{\mathrm{FI}}(h, x, y, a).$$

Whenever the maximizer is unique on $\mathrm{supp}(p_t^h)$, the local game satisfies the same strict-optimality assumption as in the main body.

By construction,

$$V^\star\big(M_{t,h}^{\mathrm{loc}}\big) = V_t^{\mathrm{FI}}(h). \tag{9}$$

I.3. Dynamic Gap Decomposition

We first compare any deterministic safe policy with the full-information benchmark.

Lemma I.4 (Bellman domination)**.** *For every deterministic safe policy π, every stage $t \in [H]$, and every public history $h \in \mathcal{H}_t$,*

$$V_t^\pi(h) \le V_t^{\mathrm{FI}}(h).$$

Proof. The proof is by backward induction on t.

The claim is trivial for $t = H + 1$ because both quantities are zero. Suppose it holds at stage $t + 1$. Fix $h \in \mathcal{H}_t$. By definition of A_t^π,

$$\begin{aligned}
V_t^\pi(h) &= \sum_{(x,y)} p_t^h(x, y)\, \mathbb{E}\big[R_t + V_{t+1}^\pi(H_{t+1}) \,\big|\, H_t = h,\ X_t = x,\ Y_t = y,\ A_t = A_t^\pi\big] \\
&\le \sum_{(x,y)} p_t^h(x, y)\, \mathbb{E}\big[R_t + V_{t+1}^{\mathrm{FI}}(H_{t+1}) \,\big|\, H_t = h,\ X_t = x,\ Y_t = y,\ A_t = A_t^\pi\big] \\
&= \sum_{(x,y)} p_t^h(x, y)\, Q_t^{\mathrm{FI}}(h, x, y, A_t^\pi) \\
&\le \sum_{(x,y)} p_t^h(x, y) \max_{a \in \mathcal{A}_t(h)} Q_t^{\mathrm{FI}}(h, x, y, a) \\
&= V_t^{\mathrm{FI}}(h).
\end{aligned}$$

This completes the induction. □

Define the dynamic gap

$$\Delta_t^\pi(h) := V_t^{\mathrm{FI}}(h) - V_t^\pi(h) \ge 0.$$

Lemma I.5 (Local-plus-future gap decomposition)**.** *Fix a deterministic safe policy π, a stage $t \in [H]$, and a public history $h \in \mathcal{H}_t$. Then*

$$\Delta_t^\pi(h) = \sum_{(x,y)} p_t^h(x,y) \underbrace{\Big[\max_{a\in\mathcal{A}_t(h)} Q_t^{\mathrm{FI}}(h,x,y,a) - Q_t^{\mathrm{FI}}(h,x,y,A_t^\pi(h,x,y))\Big]}_{=:\Gamma_t^\pi(h,x,y)} + \sum_{(x,y)} p_t^h(x,y)\,\mathbb{E}\big[\Delta_{t+1}^\pi(H_{t+1}) \,\big|\, H_t = h,\ X_t = x,\ Y_t = y,\ A_t = A_t^\pi(h,x,y)\big]\,. \tag{10}$$

In particular, both terms on the right-hand side are nonnegative.

Proof. Starting from the definitions of $V_t^{\mathrm{FI}}(h)$ and $V_t^\pi(h)$,

$$\Delta_t^\pi(h) = \sum_{(x,y)} p_t^h(x,y) \max_a Q_t^{\mathrm{FI}}(h,x,y,a) - \sum_{(x,y)} p_t^h(x,y)\,\mathbb{E}\big[R_t + V_{t+1}^\pi(H_{t+1}) \,\big|\, H_t = h,\ X_t = x,\ Y_t = y,\ A_t = A_t^\pi\big]\,.$$

Add and subtract

$$\sum_{(x,y)} p_t^h(x,y)\, Q_t^{\mathrm{FI}}(h,x,y,A_t^\pi(h,x,y)).$$

Using the definition of Q_t^{FI}, we obtain

$$\Delta_t^\pi(h) = \sum_{(x,y)} p_t^h(x,y) \Big[\max_a Q_t^{\mathrm{FI}}(h,x,y,a) - Q_t^{\mathrm{FI}}(h,x,y,A_t^\pi(h,x,y))\Big] + \sum_{(x,y)} p_t^h(x,y)\,\mathbb{E}\big[V_{t+1}^{\mathrm{FI}}(H_{t+1}) - V_{t+1}^\pi(H_{t+1}) \,\big|\, H_t = h,\ X_t = x,\ Y_t = y,\ A_t = A_t^\pi(h,x,y)\big]\,.$$

The second line is exactly the conditional expectation of $\Delta_{t+1}^\pi(H_{t+1})$, which gives Equation (10). Nonnegativity of both terms is immediate from the definitions. □

The next corollary is the key bridge from sequential exact optimality to the one-shot theory.

Corollary I.6 (Zero root gap propagates along reachable histories)**.** *Let π be a deterministic safe policy such that*

$$V_1^\pi(h_1^\circ) = V_1^{\mathrm{FI}}(h_1^\circ).$$

Then for every stage $t \in [H]$ and every public history $h \in \mathcal{H}_t$ reachable under π,

$$\Delta_t^\pi(h) = 0.$$

Moreover, for every such (t,h) and every support point $(x,y) \in \mathrm{supp}(p_t^h)$,

$$A_t^\pi(h,x,y) = a_{t,h}^\star(x,y),$$

provided the local continuation game $M_{t,h}^{\mathrm{loc}}$ satisfies the strict-optimality condition.

Proof. Since

$$\Delta_1^\pi(h_1^\circ) = V_1^{\mathrm{FI}}(h_1^\circ) - V_1^\pi(h_1^\circ) = 0,$$

Theorem I.5 implies that, at h_1°, both nonnegative terms on the right-hand side of Equation (10) vanish.

In particular, the second term vanishes:

$$\mathbb{E}[\Delta_2^\pi(H_2) \,|\, H_1 = h_1^\circ,\ X_1,\ Y_1,\ A_1 = A_1^\pi] = 0 \qquad \text{almost surely.}$$

Since $\Delta_2^\pi(H_2) \geq 0$, this implies

$$\Delta_2^\pi(h') = 0$$

for every stage-2 public history h' reached with positive probability under π. Repeating the same argument inductively proves that

$$\Delta_t^\pi(h) = 0$$

for every reachable public history h.

Now fix a reachable (t, h) and suppose the local continuation game satisfies strict optimality. Since $\Delta_t^\pi(h) = 0$, the first nonnegative term in Equation (10) must vanish. Therefore, for every $(x, y) \in \mathrm{supp}(p_t^h)$,

$$\max_a Q_t^{\mathrm{FI}}(h, x, y, a) = Q_t^{\mathrm{FI}}(h, x, y, A_t^\pi(h, x, y)).$$

By strict optimality of the local continuation game,

$$A_t^\pi(h, x, y) = a_{t,h}^\star(x, y),$$

as claimed. □

I.4. Local Chromatic Lower Bound

We can now state the sequential bridge theorem.

Theorem I.7 (Sequential local lower bound)**.** *Let π be a deterministic safe policy for a finite-horizon directional POAG such that*

$$V_1^\pi(h_1^\circ) = V_1^{\mathrm{FI}}(h_1^\circ).$$

Fix a stage $t \in [H]$ and a public history $h \in \mathcal{H}_t$ reachable under π. Assume that the local continuation game

$$M_{t,h}^{\mathrm{loc}} = \big(\mathcal{X}_t(h),\, \mathcal{Y}_t(h),\, \mathcal{A}_t(h),\, p_t^h,\, u_{t,h}^{\mathrm{loc}}\big)$$

satisfies the strict-optimality condition. Define the local encoder and decoder

$$e_{t,h}^\pi(y) := e_t^\pi(h, y), \qquad d_{t,h}^\pi(x, m) := d_t^\pi(h, x, m).$$

Then:

1. *the local protocol $(e_{t,h}^\pi, d_{t,h}^\pi)$ attains the full-information value of $M_{t,h}^{\mathrm{loc}}$ exactly;*
2. *its message classes are independent sets of the local safe-confusability graph $G_{t,h}^{\mathrm{loc}}$;*
3. *the number of used messages at history h satisfies*

$$|\{e_t^\pi(h, y) : y \in \mathcal{Y}_t(h)\}| \geq \chi\big(G_{t,h}^{\mathrm{loc}}\big)\,;$$

equivalently, the local explicit communication at history h obeys

$$\lceil \log_2 |\{e_t^\pi(h, y) : y \in \mathcal{Y}_t(h)\}| \rceil \geq \mathrm{SRC}_{\mathrm{FL}}\big(M_{t,h}^{\mathrm{loc}}\big)\,.$$

Proof. Fix a reachable history h. By Theorem I.6, for every

$$(x, y) \in \mathrm{supp}(p_t^h),$$

the action induced by π at history h equals the unique locally optimal action:

$$d_t^\pi(h, x, e_t^\pi(h, y)) = A_t^\pi(h, x, y) = a_{t,h}^\star(x, y).$$

Hence the local protocol $(e_{t,h}^\pi, d_{t,h}^\pi)$ realizes the unique full-information optimal action of $M_{t,h}^{\mathrm{loc}}$ at every support point. Therefore

$$V_{M_{t,h}^{\mathrm{loc}}}(e_{t,h}^\pi, d_{t,h}^\pi) = V^\star(M_{t,h}^{\mathrm{loc}}),$$

which proves part (1).

Part (2) now follows from the one-shot fixed-length theorem proved earlier: since the local protocol attains full-information value exactly, each of its message classes must be an independent set of the local safe-confusability graph.

For part (3), let

$$\mathcal{M}_{t,h}^{\text{used}} := \{e_t^{\pi}(h, y) : y \in \mathcal{Y}_t(h)\}.$$

By part (2), the partition of $\mathcal{Y}_t(h)$ induced by the map $e_t^{\pi}(h, \cdot)$ is a proper coloring of $G_{t,h}^{\text{loc}}$ by the colors in $\mathcal{M}_{t,h}^{\text{used}}$. Hence

$$|\mathcal{M}_{t,h}^{\text{used}}| \geq \chi(G_{t,h}^{\text{loc}}).$$

Applying the one-shot characterization to the local game gives

$$\text{SRC}_{\text{FL}}(M_{t,h}^{\text{loc}}) = \lceil \log_2 \chi(G_{t,h}^{\text{loc}}) \rceil ,$$

which implies the stated bit lower bound. □

Remark I.8 (What the theorem does and does not say). Theorem I.7 is a local necessary condition, not a full sequential characterization. It does not claim that local chromatic lower bounds add across time, nor that satisfying them is sufficient for exact sequential full-information attainment. What it does show is that the one-shot safe-confusability graph is the correct obstruction at any reachable stage of a finite-horizon POAG whenever exact globally full-information-optimal safe coordination is achieved.

J. Experimental Details

This appendix specifies the exact computational procedures underlying Section 7. Every quantity in the main paper is obtained either by exact combinatorial optimization, exact mixed-integer linear optimization, or deterministic numerical optimization of a finite variational problem. We do *not* use learned policies, policy-gradient methods, Monte Carlo training, or stochastic approximation in the main experiments. All logarithms are base two, all reported rates are in bits, and all graphs are finite, simple, and undirected.

J.1. Common Computational Conventions

The fixed-length experiments operate on finite one-shot games with explicit message alphabets of size $q = 2^B$, where B is the bit budget. For every such instance, the theorem prediction

$$B_{\text{thm}}^{\star} = \lceil \log_2 \chi(G_M) \rceil$$

is computed from the recovered safe-confusability graph G_M using an exact branch-and-bound coloring routine. The direct protocol-design quantity $B_{\text{MILP}}^{\star}$ is computed independently from the game itself by solving a family of exact mixed-integer linear programs and selecting the smallest budget whose optimum matches the full-information value.

Unless otherwise stated, theorem/solver agreement is declared exact when the mixed-integer solver terminates with an optimality certificate and the absolute difference between the reported numerical optimum and the theoretical value is at most numerical precision. In the asymptotic-rate experiments, all independent sets are enumerated exactly, and the conditional graph-entropy optimization problem is solved deterministically to a relative improvement tolerance of 10^{-12} with a maximum of 2×10^4 fixed-point iterations. The reported reference/solver discrepancies in Table 6 are at machine precision.

The random-graph sweep in Experiment 1 uses Erdős–Rényi instances $G(n, 1/2)$ with $n \in \{6, 8, 10, 12\}$ and twenty fixed seeds per value of n. All other experiments are fully deterministic once their parameter grids are specified.

J.2. Experiment 1: Exact Threshold Recovery on Graph-Realized Instances

For Experiment 1, each input graph $G = (V, E)$ is converted into the graph realization game M_G from Theorem 5.1. The receiver observation alphabet is the tagged union

$$\{x_v : v \in V\} \cup \{x_e : e \in E\},$$

Table 4. Random-graph summary for exact threshold recovery on Erdős–Rényi instances. The threshold-count columns show the empirical distribution of theorem/MILP thresholds across random graphs of size n.

n	Num. graphs	$\#(B^\star = 1)$	$\#(B^\star = 2)$	$\#(B^\star = 3)$	Recovery rate	Agreement rate	Median total time (s)
6	20	5	15	0	1.000	1.000	0.037
8	20	1	19	0	1.000	1.000	0.094
10	20	0	13	7	1.000	1.000	0.467
12	20	0	14	6	1.000	1.000	0.612

the sender symbols and receiver actions are both identified with V, the support consists of the pairs

$$(x_v, v) \quad \text{and} \quad (x_{\{u,v\}}, u),\ (x_{\{u,v\}}, v) \qquad \text{for } \{u, v\} \in E,$$

and the payoff is

$$u(a, x, y) = \mathbf{1}\{a = y\}.$$

The recovered graph G_{M_G} is constructed *directly from the game* by declaring y and y' adjacent whenever some receiver observation x supports both sender symbols with positive probability. The graph-realization theorem predicts that $G_{M_G} = G$, and the experiment verifies this identity exactly.

For a fixed explicit-message budget B and hence a fixed message cardinality $q = 2^B$, the direct protocol-design optimum is obtained from the following mixed-integer linear program. Let $S_M \subseteq \mathcal{X} \times \mathcal{Y}$ denote the support of the realized game. Since the reward equals $\mathbf{1}\{a = y\}$, it suffices to introduce:

- binary variables $z_{y,m}$ indicating whether sender symbol y is assigned to message $m \in [q]$;
- binary variables $d_{x,m,a}$ indicating whether the decoder outputs action a after observing (x, m);
- continuous variables $s_{x,y,m}$ linearizing the product $z_{y,m} d_{x,m,y}$ on the support.

The exact direct protocol value is the optimum of

$$\begin{aligned}
\max \quad & \sum_{(x,y)\in S_M} p(x, y) \sum_{m=1}^{q} s_{x,y,m} \\
\text{s.t.} \quad & \sum_{m=1}^{q} z_{y,m} = 1 \qquad \forall y \in \mathcal{Y}, \\
& \sum_{a\in\mathcal{A}(x)} d_{x,m,a} = 1 \qquad \forall x \in \mathcal{X},\ \forall m \in [q], \\
& s_{x,y,m} \le z_{y,m}, \qquad s_{x,y,m} \le d_{x,m,y}, \qquad s_{x,y,m} \ge z_{y,m} + d_{x,m,y} - 1, \\
& z_{y,m} \in \{0, 1\}, \qquad d_{x,m,a} \in \{0, 1\}, \qquad 0 \le s_{x,y,m} \le 1.
\end{aligned} \tag{11}$$

Here $\mathcal{A}(x)$ denotes the feasible receiver actions at observation x, which in the graph-realization family is exactly the set of sender symbols supported by x. The threshold $B^\star_{\text{MILP}}$ is then the smallest B for which the optimum of Equation (11) equals $V^\star(M_G) = 1$.

The named-benchmark table appears in the main paper. The appendix-only random-graph breakdown is reported in Table 4.

J.3. Experiment 2: Missing-Bit Law and Covert Collapse

For Experiment 2 we use the family M_k from Theorem G.1. Here

$$\mathcal{X}_k = \{x_0\}, \qquad \mathcal{Y}_k = \mathcal{A}_k = [2^k], \qquad u_k(a, x_0, y) = \mathbf{1}\{a = y\},$$

and the sender symbol Y is uniform on $[2^k]$. The theorem gives the exact safe explicit value law

$$V^\star_{\text{safe}}(B) = \min\{1, 2^{B-k}\}.$$

Table 5. Exact validation of the missing-bit law via direct protocol-design MILP. For each $k \leq 8$, the direct optimizer agrees exactly with the theorem value at every budget $B \in \{0, \ldots, k\}$. The zero-bit covert value is 1 for all k.

k	2^k	Budgets checked	$\max_B \lvert V^{\mathrm{MILP}}_{k,B} - V^{\mathrm{thm}}_{k,B} \rvert$	$\mathrm{SRC}_{\mathrm{FL}}(M_k)$	$V^\star_{\mathrm{safe}}(0)$	$V^\star_{\mathrm{safe}}(k-1)$	$V^\star_{\mathrm{covert}}(0)$
2	4	$0, \ldots, 2$	0.0e+00	2	2^{-2}	$\frac{1}{2}$	1
3	8	$0, \ldots, 3$	0.0e+00	3	2^{-3}	$\frac{1}{2}$	1
4	16	$0, \ldots, 4$	0.0e+00	4	2^{-4}	$\frac{1}{2}$	1
5	32	$0, \ldots, 5$	0.0e+00	5	2^{-5}	$\frac{1}{2}$	1
6	64	$0, \ldots, 6$	0.0e+00	6	2^{-6}	$\frac{1}{2}$	1
7	128	$0, \ldots, 7$	0.0e+00	7	2^{-7}	$\frac{1}{2}$	1
8	256	$0, \ldots, 8$	0.0e+00	8	2^{-8}	$\frac{1}{2}$	1

The main-text figure plots this law for $k \in \{4, 8, 12\}$ and overlays exact direct-MILP validation points for the subset $k \leq 8$.

Because the receiver observation is constant, the direct protocol-design problem admits a smaller exact mixed-integer formulation than Equation (11). Fix k and budget B, and set

$$N := 2^k, \qquad q := 2^B.$$

For a partition of the N sender symbols into q message classes, the decoder can recover at most one sender symbol per nonempty class. Therefore the exact optimum is obtained from the binary program

$$\begin{aligned} \max \quad & \frac{1}{N} \sum_{m=1}^{q} t_m \\ \text{s.t.} \quad & \sum_{m=1}^{q} z_{i,m} = 1 \qquad \forall i \in [N], \\ & z_{i,m} \leq t_m \qquad \forall i \in [N],\ \forall m \in [q], \\ & t_m \leq \sum_{i=1}^{N} z_{i,m} \qquad \forall m \in [q], \\ & z_{i,m} \in \{0,1\}, \qquad t_m \in \{0,1\}. \end{aligned} \tag{12}$$

Here $z_{i,m} = 1$ indicates that sender symbol i is assigned to message m, and $t_m = 1$ indicates that class m is nonempty and can therefore contribute at most one correct recovery. This program is solved for every

$$k \in \{2, 3, \ldots, 8\}, \qquad B \in \{0, 1, \ldots, k\},$$

yielding the exact validation table in Table 5. All reported discrepancies between the theorem law and the direct optimizer are zero to numerical precision.

The covert-collapse point in the main-text figure is not estimated numerically. It is computed exactly from the theorem construction:

$$V^\star_{\mathrm{covert}}(0) = 1.$$

Likewise, the nuisance-bit decision values

$$v_{\mathrm{guess}}(Q_0) = 1, \qquad v_{\mathrm{guess}}(Q_m) = \tfrac{1}{2},$$

are computed analytically from the binary decision problem used in the proof of strict Blackwell inferiority; cf. Theorem B.12 and Blackwell (1953).

J.4. Experiment 3: Asymptotic Rate Computation

Experiment 3 validates the asymptotic characterization

$$\mathrm{SRC}_\infty(M) = H_G(Y \mid X)$$

by exact independent-set enumeration and deterministic optimization of the conditional graph-entropy variational problem from Körner (1973); Orlitsky & Roche (2001). For a graph G on the sender alphabet $\mathcal{Y}$, let $\mathcal{I}(G)$ denote the family of nonempty independent sets. The experiment solves

$$\begin{aligned} H_G(Y \mid X) = \min_{q(\cdot|y)} \quad & \sum_{x,y,W} p(x,y)\, q(W \mid y)\, \log_2 \frac{q(W \mid y)}{\sum_{y'} p(y' \mid x)\, q(W \mid y')} \\ \text{s.t.} \quad & q(W \mid y) \geq 0, \\ & \sum_{W \ni y} q(W \mid y) = 1 \qquad \forall y \in \mathcal{Y}, \\ & q(W \mid y) = 0 \qquad \text{whenever } y \notin W. \end{aligned} \tag{13}$$

Because all graphs used in the paper are small, every nonempty independent set is enumerated exactly. The optimization is then carried out by a deterministic Blahut–Arimoto-style fixed-point map derived from the KKT conditions:

$$q_{\text{new}}(W \mid y) \propto \exp\left(\sum_x p(x \mid y) \log r_x(W) \right), \qquad r_x(W) := \sum_{y'} p(y' \mid x) q(W \mid y'). \tag{14}$$

The solver terminates when the relative improvement in the objective is at most 10^{-12}, or after 2×10^4 iterations if that criterion is reached first. In all reported instances, convergence occurs well before the iteration limit.

We use two reference families.

Clique sanity family. The first family takes

$$G = K_{16},$$

a constant receiver observation X, and a Zipf-distributed sender symbol

$$\mathbb{P}(Y = i) \propto (i+1)^{-\alpha}, \qquad \alpha \in \{0, 0.5, 1, 1.5, 2\}.$$

Since every nonempty independent set of K_{16} is a singleton, one has

$$H_G(Y \mid X) = H(Y)$$

exactly. The appendix-only sanity figure in Figure 3 shows that the conditional graph-entropy solver matches this closed-form value to machine precision, while the fixed-length baseline remains

$$\log_2 \chi(K_{16}) = 4.$$

Weighted graph-realization family. The second family modifies the graph-realization construction while preserving its support graph. For a graph $G = (V, E)$ and a parameter $\rho > 0$, define

$$p_\rho(x_v, v) = \frac{\rho}{Z_\rho}, \qquad p_\rho(x_{\{u,v\}}, u) = p_\rho(x_{\{u,v\}}, v) = \frac{1}{Z_\rho}, \qquad Z_\rho := \rho|V| + 2|E|.$$

Because the support is unchanged, the safe-confusability graph remains exactly G. For this family the asymptotic rate also has a closed form:

$$H_G(Y \mid X) = \mathbb{P}_\rho\{X \text{ is an edge observation}\} \cdot 1 = \frac{2|E|}{\rho|V| + 2|E|}. \tag{15}$$

The main paper reports the ρ-sweep for C_5 and Petersen on the grid

$$\rho \in \{0.5, 1, 2, 5, 10, 20, 50\}.$$

The appendix table Table 6 reports both the solver value and the closed-form reference from Equation (15); the discrepancy is again at machine precision.

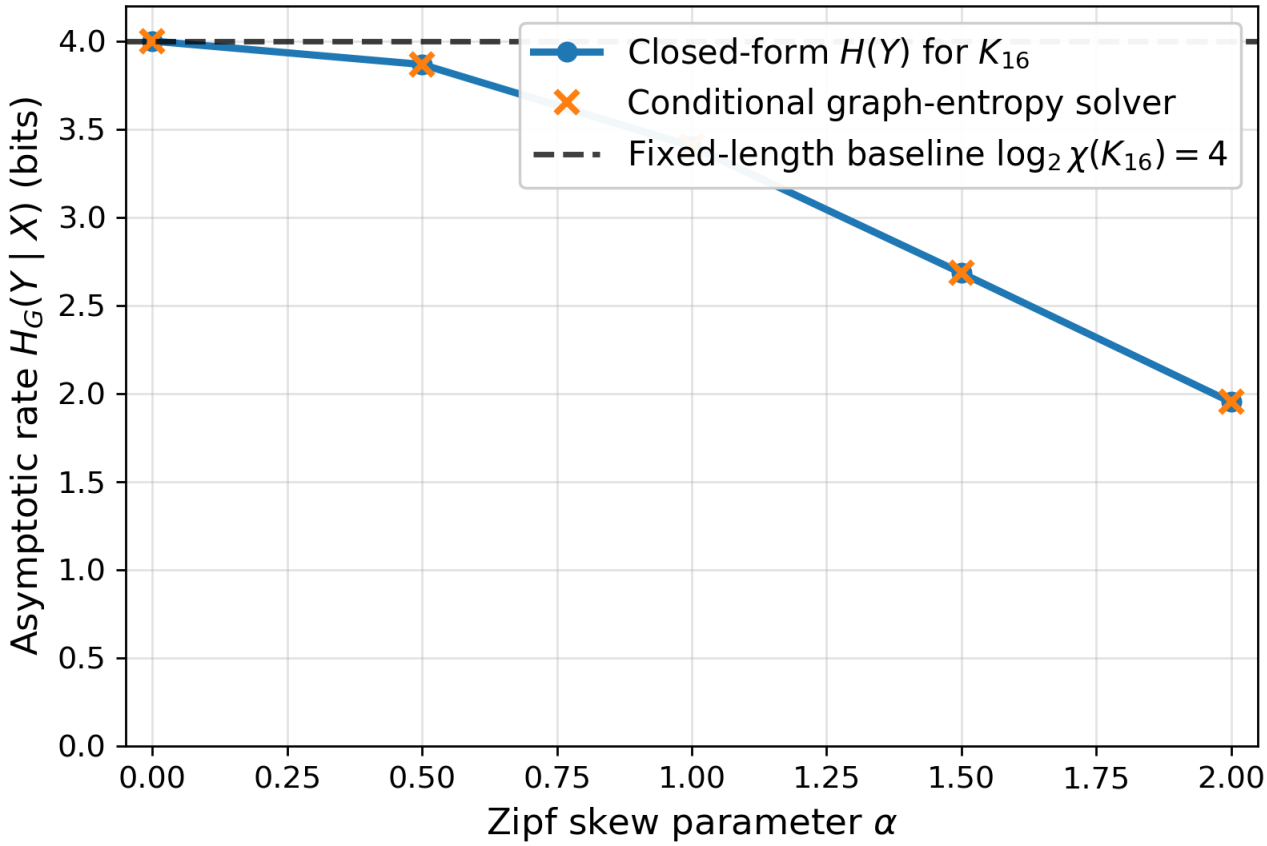


Figure 3. Closed-form clique sanity check for the conditional graph-entropy solver. For $G = K_{16}$ and constant X, the asymptotic rate equals $H(Y)$. The numerical solver matches the closed-form entropy across Zipf skew parameters α, while the fixed-length baseline remains $\log_2 \chi(K_{16}) = 4$.

J.5. Appendix-Only Sequential Bridge Experiment

The final appendix-only experiment validates the local lower-bound theorem of Theorem I.7 on a fully explicit two-stage construction. Stage 1 is public and payoff-free: it reveals one of three reachable public histories

$$h_{K_{3,3}}, \qquad h_{C_5}, \qquad h_{K_5},$$

with probabilities

$$w_{K_{3,3}} = 0.40, \qquad w_{C_5} = 0.35, \qquad w_{K_5} = 0.25.$$

Conditional on a realized public history h_G, stage 2 instantiates exactly the graph-realization local continuation game associated with the graph

$$G \in \{K_{3,3},\, C_5,\, K_5\}.$$

Thus the local safe-confusability graph is known analytically, and the theorem predicts the local message lower bound

$$q_{\mathrm{lb}}(h_G) = \chi(G).$$

For each reachable history, the exact minimum number of local messages required to attain the full-information continuation value is computed by the same graph-realization MILP used in Experiment 1. The resulting comparison appears in Figure 4 and Table 7. In every reachable history, the theorem lower bound is tight:

$$q^{\star}_{\mathrm{MILP}}(h_G) = \chi(G).$$

Table 6. Selected instances for asymptotic safe revelation. The clique rows serve as closed-form sanity checks, where $H_G(Y \mid X) = H(Y)$ because $G = K_{16}$ and X is constant. The graph-realization rows show how informative receiver side information reduces the asymptotic rate far below the fixed-length baseline $\log_2 \chi(G)$ while preserving the same conflict graph.

Instance	Param.	$\chi(G)$	$\log_2 \chi(G)$	$H_G(Y \mid X)$	Ref. value	Abs. err.	Gain
K_{16}	$\alpha = 0.0$	16	4.000	4.000	4.000	0.0e+00	0.0%
K_{16}	$\alpha = 1.5$	16	4.000	2.684	2.684	4.4e-16	32.9%
C_5	$\rho = 1.0$	3	1.585	0.667	0.667	1.1e-16	57.9%
C_5	$\rho = 10.0$	3	1.585	0.167	0.167	2.8e-17	89.5%
Petersen	$\rho = 1.0$	3	1.585	0.750	0.750	1.1e-16	52.7%
Petersen	$\rho = 10.0$	3	1.585	0.231	0.231	5.6e-17	85.4%

Table 7. Sequential local lower-bound validation on reachable stage-2 histories of the two-stage POAG. At each reachable public history h, the local continuation game is a graph-realization instance $M^{\mathrm{loc}}_{t,h}$. The theorem lower bound $q_{\mathrm{lb}} = \chi(G^{\mathrm{loc}}_{t,h})$ matches the exact minimum number of local messages required to attain the full-information continuation value.

History	w_h	Graph	$\lvert V\rvert$	$\lvert E\rvert$	$\lvert E(G^{\mathrm{loc}}) \triangle E(G)\rvert$	$\chi(G^{\mathrm{loc}})$	B_{lb}	$q^\star_{\mathrm{MILP}}$	$V_{\mathrm{loc}}(B_{\mathrm{lb}} - 1)$
$h_{K_{3,3}}$	0.40	$K_{3,3}$	6	9	0	2	1	2	0.625
h_{C_5}	0.35	C_5	5	5	0	3	2	3	0.933
h_{K_5}	0.25	K_5	5	10	0	5	3	5	0.960

We also study a *uniform local bit budget* B applied simultaneously at all reachable stage-2 histories. Because stage 1 is payoff-free and fully revealed, the global two-stage safe value is exactly the weighted average of the local budget-constrained optima:

$$V_{\mathrm{seq}}(B) = \sum_{G \in \{K_{3,3}, C_5, K_5\}} w_G \, V_G(B). \tag{16}$$

Since every history has strictly positive probability, exact full-information sequential performance is attainable if and only if the uniform budget reaches the largest reachable local lower bound in bits, namely

$$\max_G \lceil \log_2 \chi(G) \rceil = 3.$$

This is confirmed numerically by Figure 5 and Table 8.

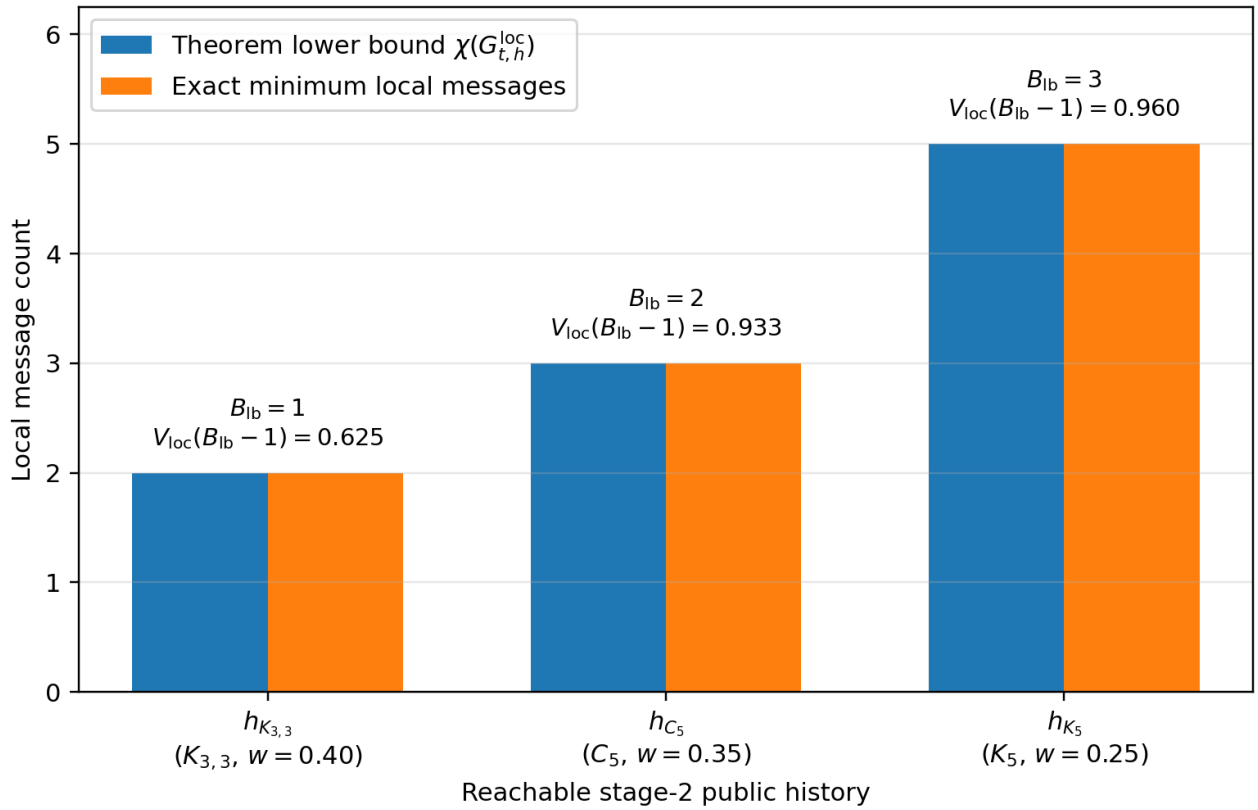


Figure 4. Sequential local lower-bound validation on reachable stage-2 public histories of the constructed two-stage POAG. At each reachable history h, the local continuation game is a graph-realization instance, and the theorem lower bound $\chi(G^{\mathrm{loc}}_{t,h})$ matches the exact minimum number of local messages required to attain the full-information continuation value.

Table 8. Global two-stage safe value under a uniform local bit budget B applied at all reachable stage-2 histories. Because every reachable history has positive probability, exact full-information sequential performance is attained only when B is at least the largest local lower-bound in bits.

B	Global safe value	Global gap	Exact?
0	0.633	0.367	No
1	0.937	0.063	No
2	0.990	0.010	No
3	1.000	0.000	Yes

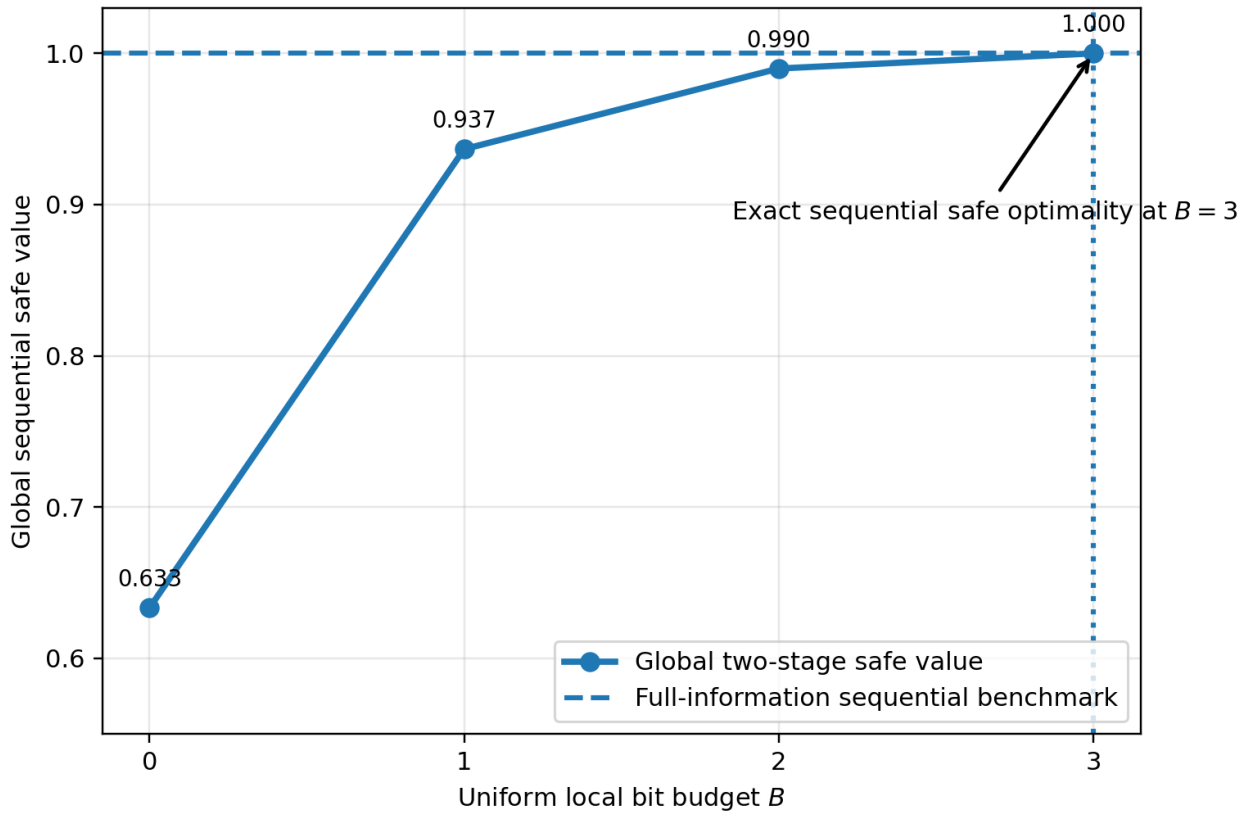


Figure 5. Global two-stage safe value under a uniform local bit budget B applied at all reachable stage-2 histories. Exact full-information sequential performance is attained only when the uniform local budget reaches the largest reachable local lower bound, here $B = 3$.