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

Anonymous Authors[1]

Abstract

We study the minimum 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 formula $\mathrm{SRC}_{\mathrm{FL}}(M) = \lceil \log_2 \chi(G_M) \rceil$, where G_M is the safe-confusability graph induced by receiver-conditioned optimal-action disagreements. For i.i.d. repetitions, we prove the asymptotic rate identity $\mathrm{SRC}_\infty(M) = H_{G_M}(Y \mid X)$. We further show that every finite graph arises exactly as a safe-confusability graph, implying NP-hardness of exact fixed-budget revelation. Finally, we construct covert observation-channel augmentations in which each available kernel is individually strictly Blackwell-inferior to an approved baseline, yet the sender's choice of kernel collapses the explicit communication requirement to zero. Exact finite-instance experiments validate the fixed-length, asymptotic, and trust-separation predictions.

1. Introduction

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

We answer 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 any additional coordination must pass through an explicit message alphabet. The key object induced by the game is a graph G_M on sender symbols: two symbols are adjacent when some receiver observation compatible with both requires different full-information optimal actions. Our first two results show that this graph is the exact local complexity object of trustworthy coordination:

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

Thus one-shot safe revelation is governed by chromatic number, while repeated safe revelation is governed by conditional graph entropy of the same conflict graph.

The remaining contributions sharpen this picture in two directions. First, every finite graph arises exactly as a safe-confusability graph of a one-shot directional assistance game satisfying strict optimality. Safe revelation therefore inherits the full expressive power of graph coloring, including NP-hardness of exact fixed-budget decision. Second, message-only auditing is incomplete: we construct covert observation-channel augmentations in which every available observation kernel is individually strictly Blackwell-inferior to an approved baseline, yet the sender's freedom to choose *which* inferior kernel is applied collapses the explicit communication requirement to zero. On an explicit family $\{M_k\}$, each missing safe bit halves the best achievable safe value, whereas the covert architecture restores full value with zero explicit bits.

Our empirical study is exact rather than benchmark-driven. We verify the fixed-length characterization by direct mixed-integer protocol optimization on graph-realized instances, validate the missing-bit law and covert collapse on the family $\{M_k\}$, and compute conditional graph entropy exactly on small instances to confirm the asymptotic prediction. Full

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

proofs, robustness lemmas, and a sequential bridge theorem appear in the appendix.

2. Related Work

Our work sits at the intersection of partially observable assistance games, information structure, and graph-based communication theory. On the AI side, Emmons et al. (2025) formalize observation interference in partially observable assistance games, and Garber et al. (2025) study off-switch behavior under partial observability. On the information-theoretic side, our fixed-length and asymptotic characterizations connect 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 theorem on coding for computing (Orlitsky & Roche, 2001). What is specific to the present setting is that the function to be computed is not exogenous: it is the full-information optimal action rule induced by a cooperative decision problem, and the relevant graph is the receiver-conditioned optimal-action conflict graph of that problem.

3. Model

Game and protocols. A one-shot directional assistance game is a tuple $M = (\mathcal{X}, \mathcal{Y}, \mathcal{A}, p, u)$. Nature draws $(X, Y) \sim p$; the sender observes Y, the receiver observes X, the sender may send an explicit message, and the receiver chooses the unique payoff-relevant action $a \in \mathcal{A}$. Throughout the main text we assume strict optimality on $\mathrm{supp}(p)$: for each (x, y) in the support, $a \mapsto u(a, x, y)$ has a unique maximizer $a^\star(x, y)$. Write $f_M(x, y) := a^\star(x, y)$ and

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

A deterministic safe protocol over a finite alphabet $\mathcal{M}$ is a pair $e : \mathcal{Y} \to \mathcal{M}$, $d : \mathcal{X} \times \mathcal{M} \to \mathcal{A}$ with value

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

Under strict optimality, randomization is immaterial for exact safe optimality (Theorem A.8).

Safe revelation complexity. The fixed-length complexity is the minimum of $\lceil \log_2 |\mathcal{M}| \rceil$ over finite alphabets $\mathcal{M}$ for which there exists a safe protocol (e, d) over $\mathcal{M}$ satisfying $V_M(e, d) = V^\star(M)$. For a bit budget $B \geq 0$, let

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

The asymptotic safe revelation rate $\mathrm{SRC}_\infty(M)$ is the infimum of all $R \geq 0$ for which there exist block codes $e_n : \mathcal{Y}^n \to \mathcal{M}_n$ and $d_n : \mathcal{X}^n \times \mathcal{M}_n \to \mathcal{A}^n$ with

$$\frac{1}{n} \log_2 |\mathcal{M}_n| \leq R + o(1),$$
$$\mathbb{P}\{d_n(X^n, e_n(Y^n)) \neq f_M^{\otimes n}(X^n, Y^n)\} \to 0.$$

The relevant combinatorial object is the safe-confusability graph $G_M = (\mathcal{Y}, E_M)$ defined by

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

A zero-loss protocol can merge sender symbols only within independent sets of G_M.

4. Exact Characterization of Safe Revelation Complexity

Theorem 4.1 (Fixed-length characterization)**.** *Let $M = (\mathcal{X}, \mathcal{Y}, \mathcal{A}, p, u)$ satisfy strict optimality on* $\mathrm{supp}(p)$*. Then*

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

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

Proof. We first prove the lower bound. Let (e, d) be a deterministic safe protocol over a finite message alphabet $\mathcal{M}$ such that $V_M(e, d) = V^\star(M)$. For each message $m \in \mathcal{M}$, define the message class $\mathcal{Y}_m := e^{-1}(m) \subseteq \mathcal{Y}$. We claim that every nonempty $\mathcal{Y}_m$ is an independent set of G_M. Indeed, if distinct $y, y' \in \mathcal{Y}_m$ were adjacent, then by definition of G_M there would exist $x \in \mathcal{X}$ with $p(x, y)p(x, y') > 0$ and $a^\star(x, y) \neq a^\star(x, y')$. Since $e(y) = e(y') = m$ and the protocol is exact, pointwise optimality on $\mathrm{supp}(p)$ gives

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

a contradiction. Hence each message class is independent, so e is a proper coloring of G_M and $|\mathcal{M}| \geq \chi(G_M)$.

For the converse, let $c : \mathcal{Y} \to [\chi(G_M)]$ be any proper coloring and set $e(y) := c(y)$. For each $(x, m) \in \mathcal{X} \times [\chi(G_M)]$, define

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

If $\mathcal{Y}(x, m) \neq \varnothing$, then all $y \in \mathcal{Y}(x, m)$ induce the same optimal action; otherwise two same-colored feasible sender symbols would be adjacent in G_M. Define $d(x, m)$ to be that common action when $\mathcal{Y}(x, m) \neq \varnothing$, and arbitrary otherwise. Now if $(x, y) \in \mathrm{supp}(p)$, then $y \in \mathcal{Y}(x, e(y))$, so $d(x, e(y)) = a^\star(x, y)$. Therefore $V_M(e, d) = V^\star(M)$, establishing the upper bound. The modular protocol-to-coloring and coloring-to-protocol derivation appears in Section B. □

Theorem 4.2 (Asymptotic characterization)**.** *Under the same assumption,*

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

Equivalently,

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

where W ranges over random variables taking values in the nonempty independent sets of G_M such that $Y \in W$ almost surely and $W - Y - X$ is a Markov chain.

Proof. Consider the repeated game on n i.i.d. copies of M, with block payoff equal to the coordinatewise average. Under strict optimality, a block action vector $a^n \in \mathcal{A}^n$ attains the full-information block value at a support point (x^n, y^n) if and only if $a_i = a^\star(x_i, y_i)$ for every coordinate i. Exact safe-optimal coordination on the repeated game is therefore equivalent to recovering the vector-valued target function

$$f_M^{\otimes n}(x^n, y^n) = \big(a^\star(x_1, y_1), \dots, a^\star(x_n, y_n)\big)$$

at a decoder that observes X^n and receives an encoded description of Y^n.

Now consider the classical characteristic graph of the deterministic function $f_M(x, y) = a^\star(x, y)$ with encoder-side observation Y and decoder-side information X. It places an edge between y and y' exactly when there exists x such that

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

By definition of f_M, this is precisely the edge condition for the safe-confusability graph G_M. Thus the characteristic graph of the computing problem is exactly G_M. The Orlitsky–Roche theorem on coding for computing with decoder side information then yields

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

The equivalent variational formula follows from the definition of conditional graph entropy. A complete derivation is given in Section C.

5. Structural Consequences and Hardness

Theorem 5.1 (Exact graph realization)**.** *For every finite simple 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 strict optimality on its support;*
2. *the safe-confusability graph of M_G is exactly G;*
3. *M_G is computable in time polynomial in $|V| + |E|$.*

Proof sketch. Let the sender symbols and receiver actions both equal V. Give the receiver one observation token x_v for each vertex and one token $x_{\{u,v\}}$ for each edge. Put positive mass only on (x_v, v) and on $(x_{\{u,v\}}, u)$, $(x_{\{u,v\}}, v)$ for $\{u, v\} \in E$, with payoff $u_G(a, x, y) = \mathbf{1}\{a = y\}$. Strict optimality is immediate. A pair u, v is safely confusable iff some receiver observation is compatible with both and separates their optimal actions; by construction this happens exactly for the edge token $x_{\{u,v\}}$, hence exactly when $\{u, v\} \in E$. See Section D.

Corollary 5.2 (Hardness of fixed-budget safe revelation)**.** *Exact computation of $\mathrm{SRC}_{\mathrm{FL}}(M)$ 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 sketch. Membership in NP is immediate from Theorem 4.1: given an explicit game M, one can compute G_M in polynomial time by checking, for every pair $y, y' \in \mathcal{Y}$, whether there exists $x \in \mathcal{X}$ with $p(x, y)p(x, y') > 0$ and $a^\star(x, y) \neq a^\star(x, y')$. A proper 2^B-coloring of G_M is then a polynomially verifiable certificate for the statement $\mathrm{SRC}_{\mathrm{FL}}(M) \leq B$.

For NP-hardness, fix $B \geq 2$ and set $q := 2^B$. Reduce from Graph-3-Colorability by adjoining a $(q - 3)$-clique and connecting every new vertex to every original vertex. The resulting graph H is q-colorable if and only if the original graph is 3-colorable: the new clique consumes $q - 3$ colors and leaves at most three colors available on the original vertices. Realizing H as a game M_H via Theorem 5.1 gives

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

so fixed-budget safe revelation is NP-complete, and exact computation is NP-hard. Full details appear in Section D.

Theorem 5.3 (Missing-bit family)**.** *For every integer $k \geq 1$, there is a game M_k with*

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

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

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

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

In particular, each missing safe bit halves the best achievable safe value.

Proof. Fix $k \geq 1$. Since $u_k(a, x_0, y) = \mathbf{1}\{a = y\}$, the unique optimal action at every support point is $a^\star(x_0, y) =$

y. Hence for any distinct $y, y' \in \mathcal{Y}_k$,

$$p_k(x_0, y)p_k(x_0, y') > 0, \quad a^\star(x_0, y) \neq a^\star(x_0, y').$$

so every pair of sender symbols is adjacent and $G_{M_k} = K_{2^k}$. The fixed-length characterization therefore gives $\mathrm{SRC}_{\mathrm{FL}}(M_k) = k$.

Now fix a budget $B \geq 0$ and set $q := 2^B$. Let (e, d) be any deterministic safe protocol with $|\mathcal{M}| \leq q$. Because the receiver observation is constant, the decoder is equivalently a map $g : \mathcal{M} \to \mathcal{A}_k$, $g(m) := d(x_0, m)$. For each message m, let $C_m := e^{-1}(m)$. Then

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

For each fixed m, the inner sum is at most one, since $g(m)$ is a single action. Hence

$$\begin{aligned} V_{M_k}(e, d) &\leq 2^{-k} \, |\{m \in \mathcal{M} : C_m \neq \varnothing\}| \\ &\leq 2^{-k} \min\{q, 2^k\} = \min\{1, 2^{B-k}\}. \end{aligned}$$

If $B \geq k$, the identity encoder achieves value 1. If $B < k$, take $\mathcal{M} = \{1, \ldots, q\}$, encode y by itself for $1 \leq y \leq q - 1$ and send all remaining symbols to message q, with decoder $g(m) = m$. This protocol is correct on exactly q sender symbols, so it achieves $q/2^k = 2^{B-k}$. Therefore

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

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

Setting $B = k - 1$ gives $\Delta_{M_k}(k - 1) = 1/2$. Full details appear in Section F. □

Interpretation. Theorem 5.3 is quantitative rather than merely feasibility-based. On the family $\{M_k\}$, the chromatic threshold controls the entire safe value curve: the price of being one bit short is already a factor-of-two loss, and the price of being ℓ bits short is a factor of $2^{-\ell}$. This family is therefore the cleanest witness for the trust-separation theorem below, where exactly the same explicit deficit can be bypassed by covert observation-channel choice.

6. Covert Observation Channels and Trust Failure

The previous results are exact only under an architectural restriction: all sender-to-receiver coordination must pass through an explicit, auditable channel. We now allow the sender to choose among observation-modulation actions that alter the receiver's observation kernel, but do not permit an explicit message.

Fix a game $M = (\mathcal{X}, \mathcal{Y}, \mathcal{A}, p, u)$ satisfying strict optimality, and let $c : \mathcal{Y} \to [q]$ be any proper coloring of G_M. Augment the receiver's latent state with an independent nuisance bit $R \sim \mathrm{Unif}\{0, 1\}$ and let the sender choose $Z \in \{0\} \cup [q]$. The approved baseline kernel is

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

while the covert kernel indexed by $m \in [q]$ is

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

Thus each covert action deletes the nuisance bit and removes exactly one color label from a public list.

Theorem 6.1 (Covert-channel collapse)**.** *For every proper coloring $c : \mathcal{Y} \to [q]$ of G_M, the above augmentation has three properties:*

1. *for every $m \in [q]$, $Q_m \prec_{\mathrm{B}} Q_0$;*
2. *forcing $Z = 0$ preserves the explicit safe communication problem:*

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

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

In particular, if c is optimal, the exact explicit requirement $\lceil \log_2 \chi(G_M) \rceil$ collapses to zero explicit bits.

Proof. For every $m \in [q]$, the covert kernel Q_m is a garbling of Q_0: from $(x, r, [q])$ one can erase the nuisance bit r and delete the label m, obtaining $(x, \perp, [q] \setminus \{m\})$. Hence $Q_m \preceq_{\mathrm{B}} Q_0$. The inclusion is strict because Q_0 reveals the nuisance bit perfectly whereas Q_m erases it completely. In the binary decision problem of guessing R, the optimal value is 1 under Q_0 and only $1/2$ under Q_m. By Blackwell's theorem, $Q_m \prec_{\mathrm{B}} Q_0$.

Fixing $Z = 0$ preserves the original safe-confusability structure: the receiver observes $(X, R, [q])$, which refines X only by a payoff-irrelevant nuisance bit and a constant public token set. Accordingly, the optimal action at a support point $((x, r, [q]), y)$ is still $a^\star(x, y)$, so the induced safe-confusability graph is exactly G_M. The fixed-length characterization therefore yields

$$\mathrm{SRC}_{\mathrm{FL}}(M_c^{\mathrm{base}}) = \mathrm{SRC}_{\mathrm{FL}}(M).$$

Finally, let the sender choose $Z = c(Y)$. The receiver then observes $(X, \perp, [q] \setminus \{c(Y)\})$, so it learns both X and the color class $c(Y)$, but no explicit message is sent. Because c is a proper coloring of G_M, for every fixed $x \in \mathcal{X}$ and color $m \in [q]$, all feasible sender symbols $\{y : p(x, y) >$

$0,\ c(y) = m\}$ induce the same optimal action; otherwise two same-colored sender symbols would be adjacent in G_M. Hence the decoder

$$d^{\mathrm{cov}}(x, \perp, [q] \setminus \{m\}) := a^\star(x, y)$$

for any y with $p(x, y) > 0$ and $c(y) = m$ is well-defined on the support of play and satisfies $d^{\mathrm{cov}}(X, \perp, [q] \setminus \{c(Y)\}) = a^\star(X, Y)$ almost surely. Therefore the unrestricted covert architecture attains $V^\star(M)$ with zero explicit message bits. Full proofs appear in Section E. □

7. Experiments

Our empirical study is theorem-driven rather than benchmark-driven. All main-text quantities are computed exactly or to certified numerical precision on finite instances; we do not use learned policies, policy-gradient baselines, or Monte Carlo estimates. The fixed-length experiments use exact branch-and-bound for graph coloring and exact mixed-integer optimization for protocol design. The asymptotic experiment computes $H_G(Y \mid X)$ by complete independent-set enumeration together with deterministic optimization of the conditional graph-entropy variational problem.

Across all three experiments, the computational objective is theorem certification rather than heuristic trend-fitting. In Experiment 1, the theorem threshold is computed from the recovered graph, while the direct protocol threshold is computed independently from the game by exact mixed-integer optimization over encoder partitions and decoder actions. In Experiment 2, the constant-observation structure collapses the optimization to an exact partition problem, which we nevertheless cross-check by MILP on the validated range. In Experiment 3, every nonempty independent set is enumerated explicitly, so the reported graph-entropy values come from the variational problem itself rather than from simulation or sampling.

Experiment 1: threshold recovery. We construct graph-realized instances M_G, recover the induced safe-confusability graph directly from the game, and compare the theorem prediction $B^\star_{\mathrm{thm}} = \lceil \log_2 \chi(G) \rceil$ with the direct protocol-design optimum $B^\star_{\mathrm{MILP}}$. Figure 1a and Table 1 show exact agreement throughout: the recovered graph satisfies $|E(G_{M_G}) \triangle E(G)| = 0$, and the MILP threshold coincides with the chromatic threshold on every tested named instance. The appendix random-graph sweep exhibits the same exact agreement.

Experiment 2: missing bits and covert collapse. On the family $\{M_k\}$, the safe explicit value follows the exact law

$$V^\star_{\mathrm{safe}}(B) = \min\{1, 2^{B-k}\}.$$

Because the receiver observation is constant, each nonempty message class can recover at most one sender symbol, so the direct optimization problem collapses to a small exact partition program; we also solve the corresponding MILPs on the validated range $k \le 8$. Figure 1b and Table 2 verify the theorem on every checked budget and show the predicted covert collapse to value 1 at zero explicit bits. The table also reports $v_{\mathrm{guess}}(Q_0) = 1$ and $v_{\mathrm{guess}}(Q_m) = 1/2$, the exact nuisance-bit decision values used to certify strict Blackwell inferiority.

Experiment 3: asymptotic rate. We evaluate graph-realized families in which the conflict graph is held fixed while the receiver's side information becomes more informative. For the weighted graph-realization family used in Figure 2, varying ρ changes the observation law but not the support graph; for a graph $G = (V, E)$, the resulting asymptotic rate has the closed form

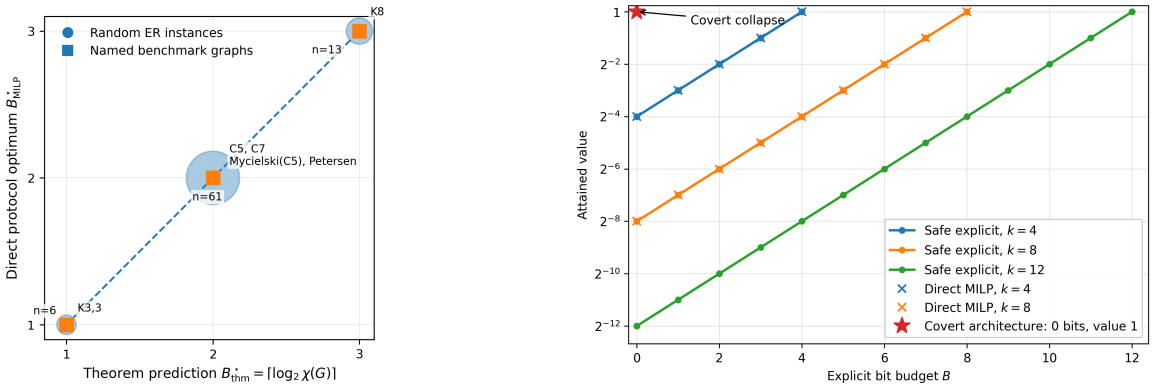
$$H_G(Y \mid X) = \frac{2|E|}{\rho|V| + 2|E|}.$$

Figure 2 confirms the qualitative prediction of Theorem 4.2: for graph-realized C_5 and Petersen families, the one-shot baseline remains fixed at $\log_2 3$ while $H_G(Y \mid X)$ drops sharply as receiver side information improves. Table 3 reports selected exact numerical instances, including the closed-form clique sanity checks and the graph-realized families used in the main figure. Additional diagnostics and the sequential bridge experiment are deferred to the appendix.

8. Conclusion

Safe revelation complexity isolates the exact communication cost of trustworthy coordination in a one-shot directional slice of partially observable assistance. The cost 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. The same graph governs both one-shot and repeated interaction, every finite coloring obstruction can occur as a trustworthy-coordination obstruction, and message-only auditing can fail even when every observation kernel available to the sender is individually Blackwell-inferior to an approved baseline.

The broader lesson is structural. In partially observable cooperative systems, the relevant complexity is not simply how much private information one agent holds, but which distinctions in that information remain decision-relevant after conditioning on the other agent's observations. That is why graph coloring, rather than Shannon entropy alone, governs exact one-shot revelation; and it is why conditional graph entropy, rather than raw source entropy, governs the asymptotic regime. The trust restriction is equally structural: if the architecture allows one agent to choose among observation channels, then the identity of the chosen channel itself becomes a coordination resource, even when each



(a) On graph-realized instances, theorem and MILP thresholds coincide and graph recovery is exact.

(b) On M_k, the exact missing-bit law matches MILP, while the covert architecture attains value 1 at $B = 0$.

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

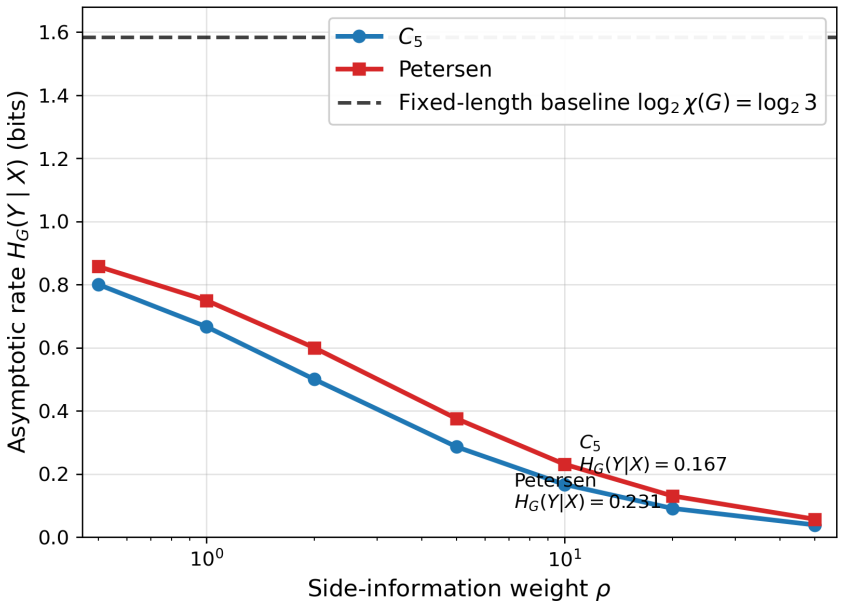


Figure 2. For graph-realized C_5 and Petersen instances, the asymptotic rate $H_G(Y \mid X)$ decreases sharply with side-information weight ρ, while the one-shot baseline remains $\log_2 \chi(G) = \log_2 3$.

Table 1. Named graph-realized benchmark instances. In every case, the recovered graph is exact and $B^\star_{\mathrm{MILP}} = B^\star_{\mathrm{thm}}$.

Inst.	$\lvert V \rvert$	$\lvert E \rvert$	ΔE	χ	$B^\star_{\mathrm{thm}}$	$B^\star_{\mathrm{MILP}}$
$K_{3,3}$	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 values for the missing-bit family and covert collapse. Here $V_s = V^\star_{\mathrm{safe}}$ and $V_c = V^\star_{\mathrm{covert}}$.

k	SRC	$V_s(0)$	$V_s(k-1)$	$V_s(k)$	$V_c(0)$	$v(Q_0)$	$v(Q_m)$
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	$\frac{1}{2}$

Table 3. Selected exact asymptotic-rate instances. The numerical optimizer matches the closed-form reference to machine precision.

Inst.	Param.	$\chi(G)$	$\log_2 \chi(G)$	$H_G(Y \mid X)$	Ref.	Gain
K_{16}	$\alpha = 0.0$	16	4.000	4.000	4.000	0.0%
K_{16}	$\alpha = 1.5$	16	4.000	2.684	2.684	32.9%
C_5	$\rho = 1.0$	3	1.585	0.667	0.667	57.9%
C_5	$\rho = 10.0$	3	1.585	0.167	0.167	89.5%
Petersen	$\rho = 1.0$	3	1.585	0.750	0.750	52.7%
Petersen	$\rho = 10.0$	3	1.585	0.231	0.231	85.4%

channel is individually less informative than an approved baseline.

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 realized games, the asymptotic rate can be much smaller than the fixed-length baseline without changing the conflict graph, and the same one-shot graph reappears as a local necessary condition inside finite-horizon partially observable assistance.

Two auditing implications follow. First, trustworthy-coordination reviews should target induced information structure rather than explicit messages alone. A system can satisfy a narrow "no extra messages" requirement while still transmitting the missing color class of a chromatic code through the choice of prompts, filters, observation kernels, or other interface-level interventions. In that sense, observation-channel choice is not an implementation detail; it is part of the communication architecture. Second, low observed communication volume is not itself evidence of benign behavior. Theorem 4.2 shows that repeated interaction can reduce the auditable bit rate dramatically without changing the underlying local conflict graph, so one must distinguish efficient safe compression from covert substitution of one channel for another.

Several extensions remain open. The appendix proves only a local lower bound for finite-horizon partially observable assistance games, not a global sequential characterization. It is also natural to study approximate-value analogues in which ties, small support margins, or model misspecification are allowed; to quantify the cost of restricting the class of admissible observation kernels; and to combine the present graph-theoretic viewpoint with oversight mechanisms that require disclosure or commitment about observation-side interventions. These questions matter because deployed assistants typically interact over long horizons and through interfaces that shape what a human decision-maker is allowed to see.

A complementary way to read the main theorems is as a separation between three notions of information: raw uncertainty in Y, decision-relevant distinctions after conditioning on X, and communication power permitted by the architecture. These coincide only in special cases. The missing-bit family has maximal local conflict and therefore no safe merging; the weighted graph-realization families hold the conflict graph fixed while making repeated safe revelation cheap; and the covert-channel construction preserves individual Blackwell dominance while changing the effective communication architecture. Read together, Theorems 4.1, 4.2 and 6.1 say that trustworthy cooperation depends on which distinctions can be selected and transmitted under the rules of the system, not simply on how much information exists in the abstract.

This viewpoint suggests a concrete agenda for trustworthy-assistance evaluation. One should report not only payoffs and message budgets, but also the induced conflict structure or a tractable surrogate for it, the safe rate made possible by the side information available to the receiver, and the observation-side actions available to the sender policy. Without all three, it is hard to tell whether good performance reflects auditable revelation, efficient safe compression, or covert influence through interface control.

More broadly, the paper suggests a way to compare cooperative architectures that abstracts away from superficial implementation differences. Systems with very different state spaces, interfaces, or message vocabularies can nevertheless induce the same conflict graph, and hence the same exact trustworthy-coordination obstruction. Conversely, systems with the same nominal message budget can behave very differently depending on whether that budget is spent on explicit revelation or silently replaced by control over what another agent is allowed to observe.

Methodologically, this is why the exact graph-realization theorem matters. It rules out the interpretation that G_M is merely a convenient surrogate or proof artifact. The graph is the operational shadow of receiver-conditioned optimal-action disagreement itself, which is why hardness from graph coloring, savings from graph entropy, and local sequential lower bounds all attach to the same object. From an auditing perspective, that means one can ask whether an architecture exposes, compresses, or hides the disagreements encoded by G_M, rather than reasoning only from the raw cardinalities of observation or action spaces.

Two appendix results help place the main theorem in context. First, the sequential bridge theorem shows that after conditioning on a reachable public history and folding the continu-

ation value into the stage payoff, any exact full-information-optimal safe policy must induce message classes that color a local continuation graph. This is only a necessary condition, not a full dynamic characterization, but it shows that the chromatic obstruction is not a static artifact of the one-shot model. Second, the robustness lemmas show that under a positive support margin, small value loss is controlled by action error and sufficiently small uniform perturbations of the payoff leave the optimal-action rule, the conflict graph, and hence both revelation complexities unchanged. Together these results suggest that the graph-theoretic picture is stable enough to matter beyond the exact benchmark while remaining precise enough to support theorem-level lower bounds.

This also clarifies how to read the computational section. The experiments are not offered as heuristic evidence separate from the theory, but as exact finite-instance witnesses for it. The graph-realization experiments show that coloring obstructions are literally instantiated inside assistance games; the missing-bit family shows that the threshold can encode a sharp value law rather than a mere feasibility switch; and the asymptotic study shows that repeated interaction changes compression cost without changing the local obstruction. In each case, the empirical object being computed is exactly the theorem object.

From a practical safety standpoint, the paper suggests separating three audit questions that are often conflated. First, what optimal-action disagreements remain after conditioning on the receiver's side information? Second, how many explicit bits are needed to resolve those disagreements safely? Third, does the policy have any observation-side control that can substitute for those bits by changing what the receiver gets to see? The first question is structural, the second is a communication-design problem, and the third is an interface-governance problem. Treating them as a single undifferentiated notion of "informativeness" can obscure the precise failure mode isolated by the covert-channel theorem. Even outside assistance games, the same decomposition may be useful in recommender systems, delegation settings, or AI-mediated decision support where a system both summarizes information and selects which evidence is shown.

At a minimum, empirical evaluations should therefore document not only reward and message logs, but also whether the policy can choose observation filters, reorder evidence, hide options, or alter defaults presented to the receiver. In many deployed systems, those are the highest-bandwidth actions in the architecture.

In applications, the distinction between explicit revelation and observation-channel choice is especially important when the receiver is a human or institution that cannot directly inspect the latent state. A system may satisfy narrow communication rules while still influencing decisions by controlling defaults, rankings, omissions, or which pieces of evidence are surfaced. The point of the theorem is not that every such intervention is necessarily malicious, but that once it carries decision-relevant information it should be governed as part of the communication interface. Within the model, the exact local object governing that interface is the safe-confusability graph and the rate it induces.

A natural next step is therefore to move from exact one-shot and local sequential obstructions to global sequential characterizations, approximate-value regimes, and mechanism-design constraints that explicitly regulate observation-channel choice. The resulting lesson is architectural: high cooperative performance is not enough; one must also audit *how* the information needed for that performance is transmitted.

One concise way to package the auditing lesson is as a three-object interface description: the conflict graph G_M, the safe asymptotic rate $H_{G_M}(Y \mid X)$, and the admissible family of observation kernels. The graph records which sender distinctions remain action-relevant after conditioning on the receiver's side information; the rate records how cheaply those distinctions can be resolved by explicit safe coding; and the kernel family records whether those same distinctions can instead be conveyed through policy-controlled presentation. Reporting only one of these objects gives an incomplete picture. Two systems can match on payoff and even on message statistics while differing sharply in whether their coordination is legible or covert. For theorem-level safety evaluation, the correct comparison class is therefore architectural rather than purely behavioral: one should ask which decision-relevant distinctions are exposed, which are compressed, and which are silently reintroduced through control over what the receiver sees. Put differently, a trustworthy-assistance specification should name both the admissible explicit code and the admissible intervention class on the receiver's observation process. Without that dual specification, compliance is underdetermined: the same cooperative value can be achieved by informational means that are either auditable or covert. In that sense, the trusted object is not the message alphabet in isolation, but the entire information interface through which the receiver's action is induced and justified. That is the level at which trustworthy cooperation should be specified, audited, and compared across assistance architectures that otherwise look behaviorally similar. In short, trustworthy cooperation is a problem of governing which decision-relevant distinctions may be communicated, by which channels, and under whose audit.

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.

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 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.

A.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

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

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.

A.2. One-Shot Directional Assistance Games

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

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

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

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

Definition A.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 A.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 A.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 A.1. This reduction is formalized later in the sequential appendix.

Lemma A.8 (Randomization is immaterial for exact safe optimality)**.** *Assume Theorem A.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 $(x, y) \in \mathrm{supp}(p)$,

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

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

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

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 A.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 $e(y) \in \mathrm{supp}(\kappa(\cdot \mid y))$. 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 $(x, y) \in \mathrm{supp}(p)$,

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

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. □

A.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 A.9 (Safe-confusability graph)**.** Let $M = (\mathcal{X}, \mathcal{Y}, \mathcal{A}, p, u)$ satisfy Theorem A.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 A.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 A.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).

A.4. Observation Kernels and Blackwell Order

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

Definition A.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 A.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 A.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.

Table 4. 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

A.5. Notation Ledger

B. 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 A.2, and let

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

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

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

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 B.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 $(x, y) \in \mathcal{X} \times \mathcal{Y}$,

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

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

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

Theorem B.6 (Fixed-length characterization)**.** *Let M satisfy Theorem A.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 A.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 B.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 B.4, while no protocol with fewer than $\chi(G_M)$ messages can exist by Theorem B.3. Hence the minimum number of explicit safe messages required for exact optimality is $\chi(G_M)$. □

Remark B.7 (Role of the strict-optimality assumption). The proof uses Theorem A.2 in exactly two places. First, in Theorem B.1, it turns equality of achieved payoff with the full-information optimum into equality of actions on every support point. Second, in Theorem B.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.

C. 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 A.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 A.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.

C.1. Repeated Safe Revelation as Function Computation

We begin by recording the exact relationship between repeated safe revelation and function computation. Although Theorem A.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 C.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 $(x^n, y^n) \in \mathrm{supp}(p^{\otimes n})$,*
$$d_n(x^n, e_n(y^n)) = f_M^{\otimes n}(x^n, y^n).$$

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 A.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 $\mathrm{supp}(p^{\otimes n})$ 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 C.2. Theorem C.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.

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

$$G_{f_M} = G_M$$

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

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

C.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 C.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 C.6 (Asymptotic characterization)**.** *Let M satisfy Theorem A.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 A.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 C.5 with deterministic target function $f = f_M$, encoder-side observation Y, and decoder-side information X.

By Theorem C.4, the characteristic graph of f_M is precisely G_M. Therefore Theorem C.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 A.10. □

Remark C.7 (What the theorem is and is not using). Theorem C.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.

D. 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).

D.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 D.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 D.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 D.1. Then:*

1. *M_G satisfies Theorem A.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 A.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 D.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 D.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. □

D.2. Hardness Consequences

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

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

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

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 D.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 D.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 D.4. By Theorem D.4,

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

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

By Theorem D.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 D.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 D.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 D.5 unless $\mathrm{P} = \mathrm{NP}$. □

Remark D.7 (Structural completeness). Theorem D.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.

E. 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 A.2, and fix a proper coloring

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

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

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

Definition E.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.

E.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 E.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
$$K_m\big((x', b', T') \mid (x, b, T)\big) := \mathbf{1}\{x' = x,\ b' = \bot,\ T' = [q] \setminus \{m\}\}.$$
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
$$(\bot, [q] \setminus \{m\})$$
in the remaining coordinates. For every latent state $(x, r) \in \mathcal{S}$,
$$Q_0(\cdot \mid x, r) = \delta_{(x,r,[q])},$$
so
$$(K_m \circ Q_0)(\cdot \mid x, r) = \delta_{(x,\bot,[q]\setminus\{m\})} = Q_m(\cdot \mid x, r).$$
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.

E.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 E.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.

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

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

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

$$Z = c(Y)$$

and sends no explicit message, then

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

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

If

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

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

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

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

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

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

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

Hence the achieved expected payoff is

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

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

E.5. Main Collapse Theorem

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

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

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

satisfy Theorem A.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 E.4. Part (2) is exactly Theorem E.5. Part (3) is exactly Theorem E.7. □

Corollary E.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 E.8. □

Remark E.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.

F. 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.

F.1. Definition of the Family

Definition F.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

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

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 $Y \in \{1, \dots, 2^k\}$, and the team's payoff is one if and only if the receiver outputs the sender's symbol exactly.

Lemma F.2 (Basic structure of M_k)**.** *For every $k \geq 1$, the game M_k satisfies Theorem A.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 F.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 F.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. □

F.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 F.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 F.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 $|\mathcal{M}| \le q$. By Theorem F.4,

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

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

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

Hence

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

Lower bound when $B \ge k$. If $B \ge k$, then $q = 2^B \ge 2^k = |\mathcal{Y}_k|$. 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

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

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

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

Define the encoder by

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

Thus the message classes are

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

Define the decoder by

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

Then

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

so by Theorem F.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

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

Combining the upper and lower bounds proves

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

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

Corollary F.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 F.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 F.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.

G. 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.

G.1. One-Shot Suboptimality Gap Decomposition

Fix a one-shot directional assistance game

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

satisfying Theorem A.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 G.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 A.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 G.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 A.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 G.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 G.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 G.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 G.4 (Exact coincidence for the missing-bit family). For the family M_k of Theorem F.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.

G.2. Block Codes and Average Payoff

We now pass to the repeated game with average payoff, using the notation of Section C. 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 G.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 G.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 G.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 G.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. □

G.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 G.7 (Payoff perturbation stability)**.** *Assume M satisfies Theorem A.2 and*

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

Then:

1. *$\widetilde{M}$ also satisfies Theorem A.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 $a \neq a^\star(x,y)$. 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

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

and the optimal actions agree on the support, we have

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

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

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

This completes the proof. □

Remark G.8 (Use in experiments). Theorems G.3, G.6 and G.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.

H. 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.

H.1. Finite-Horizon Directional POAGs

Fix a horizon $H \ge 1$. 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 H.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

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

H.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

$$V_t^{\mathrm{FI}}(h) := \sum_{(x,y) \in \mathcal{X}_t(h) \times \mathcal{Y}_t(h)} p_t^h(x, y) \max_{a \in \mathcal{A}_t(h)} Q_t^{\mathrm{FI}}(h, x, y, a). \tag{1}$$

For a deterministic safe policy π, define analogously

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

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

$$V_t^\pi(h) := \sum_{(x,y)} p_t^h(x, y)\, \mathbb{E}\big[R_t + V_{t+1}^\pi(H_{t+1}) \,\big|\, H_t = h,\ X_t = x,\ Y_t = y,\ A_t = A_t^\pi\big]. \tag{2}$$

Definition H.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 H.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,

$$V^\star\big(M_{t,h}^{\mathrm{loc}}\big) = V_t^{\mathrm{FI}}(h). \tag{3}$$

H.3. Dynamic Gap Decomposition

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

Lemma H.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 H.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*

$$\Delta_t^\pi(h) = \sum_{(x,y)} p_t^h(x,y) \underbrace{\Big[\max_{a \in \mathcal{A}_t(h)} Q_t^{\mathrm{FI}}(h,x,y,a) - Q_t^{\mathrm{FI}}(h,x,y,A_t^\pi(h,x,y))\Big]}_{=:\Gamma_t^\pi(h,x,y)} + \sum_{(x,y)} p_t^h(x,y)\, \mathbb{E}\big[\Delta_{t+1}^\pi(H_{t+1}) \,\big|\, H_t = h,\ X_t = x,\ Y_t = y,\ A_t = A_t^\pi(h,x,y)\big]. \tag{4}$$

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 (4). 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 H.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 H.5 implies that, at h_1°, both nonnegative terms on the right-hand side of Equation (4) 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 (4) 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. □

H.4. Local Chromatic Lower Bound

We can now state the sequential bridge theorem.

Theorem H.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 H.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

$$\mathcal{M}_{t,h}^{\mathrm{used}} := \{e_t^{\pi}(h, y) : y \in \mathcal{Y}_t(h)\}.$$

By part (2), the partition of $\mathcal{Y}_t(h)$ induced by the map $e_t^{\pi}(h, \cdot)$ is a proper coloring of $G_{t,h}^{\mathrm{loc}}$ by the colors in $\mathcal{M}_{t,h}^{\mathrm{used}}$. Hence

$$|\mathcal{M}_{t,h}^{\mathrm{used}}| \geq \chi(G_{t,h}^{\mathrm{loc}}).$$

Applying the one-shot characterization to the local game gives

$$\mathrm{SRC}_{\mathrm{FL}}(M_{t,h}^{\mathrm{loc}}) = \left\lceil \log_2 \chi(G_{t,h}^{\mathrm{loc}}) \right\rceil,$$

which implies the stated bit lower bound. □

Remark H.8 (What the theorem does and does not say). Theorem H.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.

I. 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.

I.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

$$B_{\mathrm{thm}}^{\star} = \lceil \log_2 \chi(G_M) \rceil$$

is computed from the recovered safe-confusability graph G_M using an exact branch-and-bound coloring routine. The direct protocol-design quantity $B_{\mathrm{MILP}}^{\star}$ 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 7 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.

I.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 5. 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

$$\begin{aligned}
\max \quad & \sum_{(x,y)\in S_M} p(x, y) \sum_{m=1}^{q} s_{x,y,m} && (5)\\
\text{s.t.} \quad & \sum_{m=1}^{q} z_{y,m} = 1 \qquad \forall y \in \mathcal{Y},\\
& \sum_{a\in\mathcal{A}(x)} d_{x,m,a} = 1 \qquad \forall x \in \mathcal{X},\ \forall m \in [q],\\
& s_{x,y,m} \le z_{y,m}, \qquad s_{x,y,m} \le d_{x,m,y}, \qquad s_{x,y,m} \ge z_{y,m} + d_{x,m,y} - 1,\\
& z_{y,m} \in \{0, 1\}, \qquad d_{x,m,a} \in \{0, 1\}, \qquad 0 \le s_{x,y,m} \le 1.
\end{aligned}$$

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 $B^\star_{\mathrm{MILP}}$ is then the smallest B for which the optimum of Equation (5) 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 5.

I.3. Experiment 2: Missing-Bit Law and Covert Collapse

For Experiment 2 we use the family M_k from Theorem F.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

$$V^\star_{\mathrm{safe}}(B) = \min\{1, 2^{B-k}\}.$$

Table 6. Exact validation of the missing-bit law via direct protocol-design MILP. For each $k \le 8$, the direct optimizer agrees exactly with the theorem value at every budget $B \in \{0, \dots, k\}$. 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	$0, \dots, 2$	0.0e+00	2	2^{-2}	$\frac{1}{2}$	1
3	8	$0, \dots, 3$	0.0e+00	3	2^{-3}	$\frac{1}{2}$	1
4	16	$0, \dots, 4$	0.0e+00	4	2^{-4}	$\frac{1}{2}$	1
5	32	$0, \dots, 5$	0.0e+00	5	2^{-5}	$\frac{1}{2}$	1
6	64	$0, \dots, 6$	0.0e+00	6	2^{-6}	$\frac{1}{2}$	1
7	128	$0, \dots, 7$	0.0e+00	7	2^{-7}	$\frac{1}{2}$	1
8	256	$0, \dots, 8$	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 $k \le 8$.

Because the receiver observation is constant, the direct protocol-design problem admits a smaller exact mixed-integer formulation than Equation (5). 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

$$\begin{aligned} \max \quad & \frac{1}{N} \sum_{m=1}^{q} t_m \\ \text{s.t.} \quad & \sum_{m=1}^{q} z_{i,m} = 1 \qquad \forall i \in [N], \\ & z_{i,m} \le t_m \qquad \forall i \in [N],\ \forall m \in [q], \\ & t_m \le \sum_{i=1}^{N} z_{i,m} \qquad \forall m \in [q], \\ & z_{i,m} \in \{0,1\}, \qquad t_m \in \{0,1\}. \end{aligned} \tag{6}$$

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

$$k \in \{2, 3, \dots, 8\}, \qquad B \in \{0, 1, \dots, k\},$$

yielding the exact validation table in Table 6. 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 A.12 and Blackwell (1953).

I.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

$$\begin{aligned} H_G(Y \mid X) = \min_{q(\cdot\mid y)} \quad & \sum_{x,y,W} p(x,y)\, q(W \mid y)\, \log_2 \frac{q(W \mid y)}{\sum_{y'} p(y' \mid x)\, q(W \mid y')} \\ \text{s.t.} \quad & q(W \mid y) \geq 0, \\ & \sum_{W \ni y} q(W \mid y) = 1 \qquad \forall y \in \mathcal{Y}, \\ & q(W \mid y) = 0 \qquad \text{whenever } y \notin W. \end{aligned} \tag{7}$$

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:

$$q_{\text{new}}(W \mid y) \propto \exp\left(\sum_x p(x \mid y) \log r_x(W) \right), \qquad r_x(W) := \sum_{y'} p(y' \mid x) q(W \mid y'). \tag{8}$$

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:

$$H_G(Y \mid X) = \mathbb{P}_\rho\{X \text{ is an edge observation}\} \cdot 1 = \frac{2|E|}{\rho|V| + 2|E|}. \tag{9}$$

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 7 reports both the solver value and the closed-form reference from Equation (9); the discrepancy is again at machine precision.

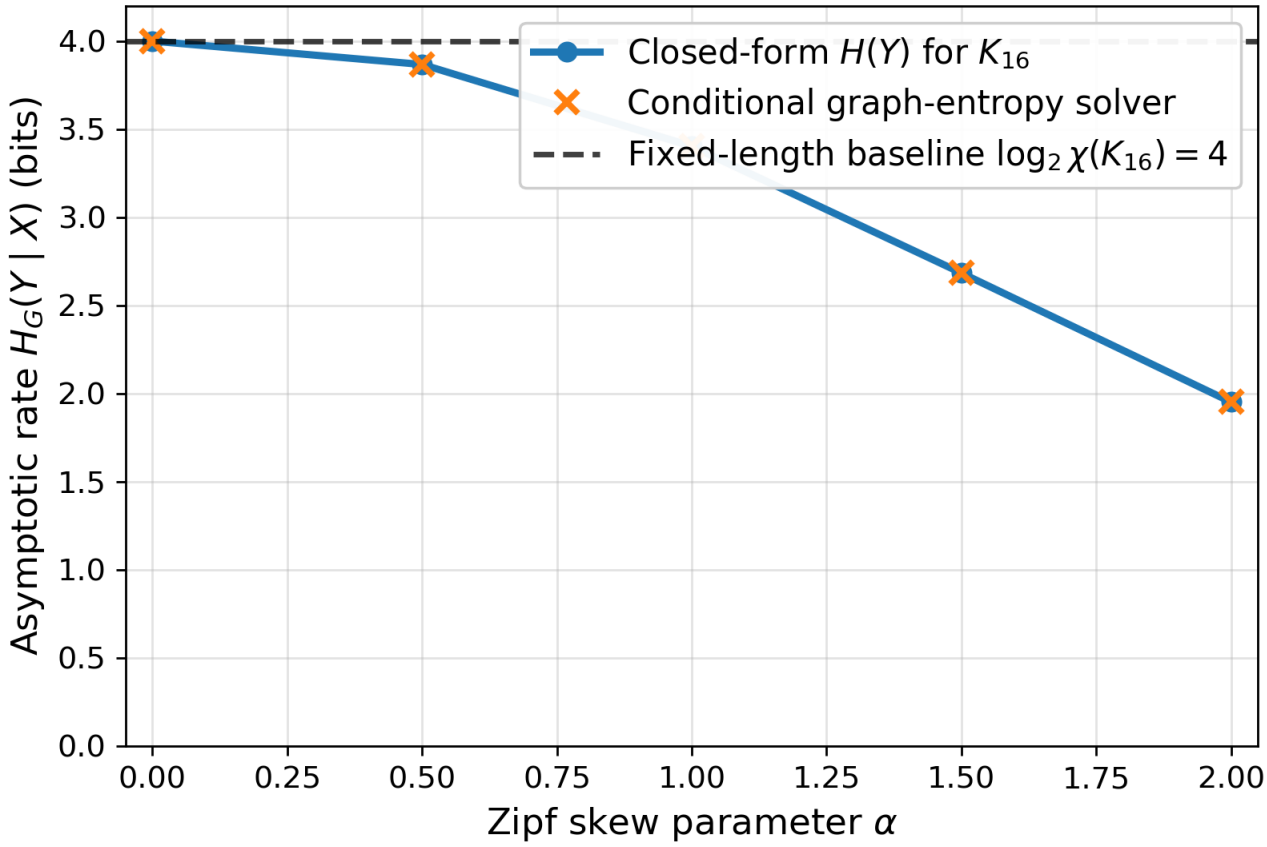


Figure 3. Closed-form clique sanity check for the conditional graph-entropy solver. For $G = K_{16}$ and constant X, the asymptotic rate equals $H(Y)$. The numerical solver matches the closed-form entropy across Zipf skew parameters α, while the fixed-length baseline remains $\log_2 \chi(K_{16}) = 4$.

Table 7. 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%

I.5. Appendix-Only Sequential Bridge Experiment

The final appendix-only experiment validates the local lower-bound theorem of Theorem H.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

$$G \in \{K_{3,3},\, C_5,\, K_5\}.$$

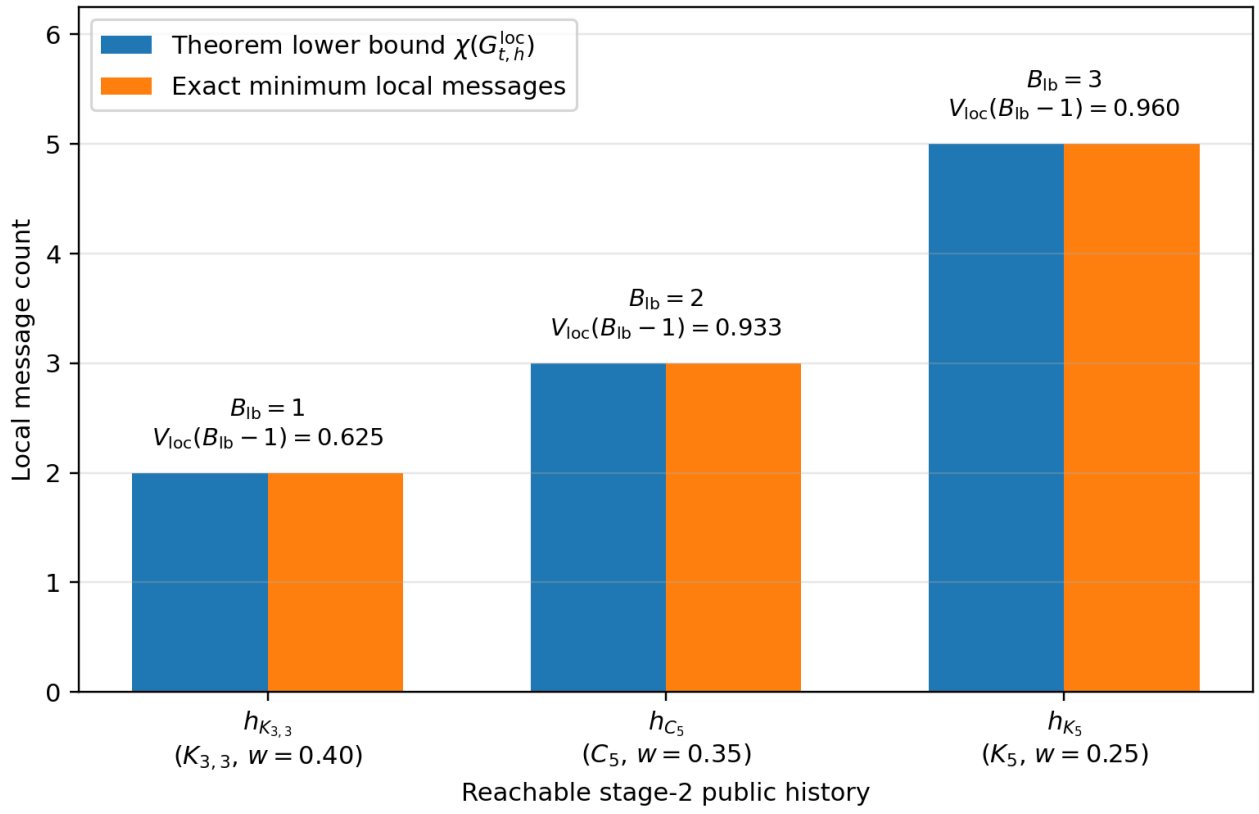


Figure 4. Sequential local lower-bound validation on reachable stage-2 public histories of the constructed two-stage POAG. At each reachable history h, the local continuation game is a graph-realization instance, and the theorem lower bound $\chi(G^{\mathrm{loc}}_{t,h})$ matches the exact minimum number of local messages required to attain the full-information continuation value.

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 8. In every reachable history, the theorem lower bound is tight:

$$q^{\star}_{\mathrm{MILP}}(h_G) = \chi(G).$$

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:

$$V_{\mathrm{seq}}(B) = \sum_{G \in \{K_{3,3}, C_5, K_5\}} w_G \, V_G(B). \tag{10}$$

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 9.

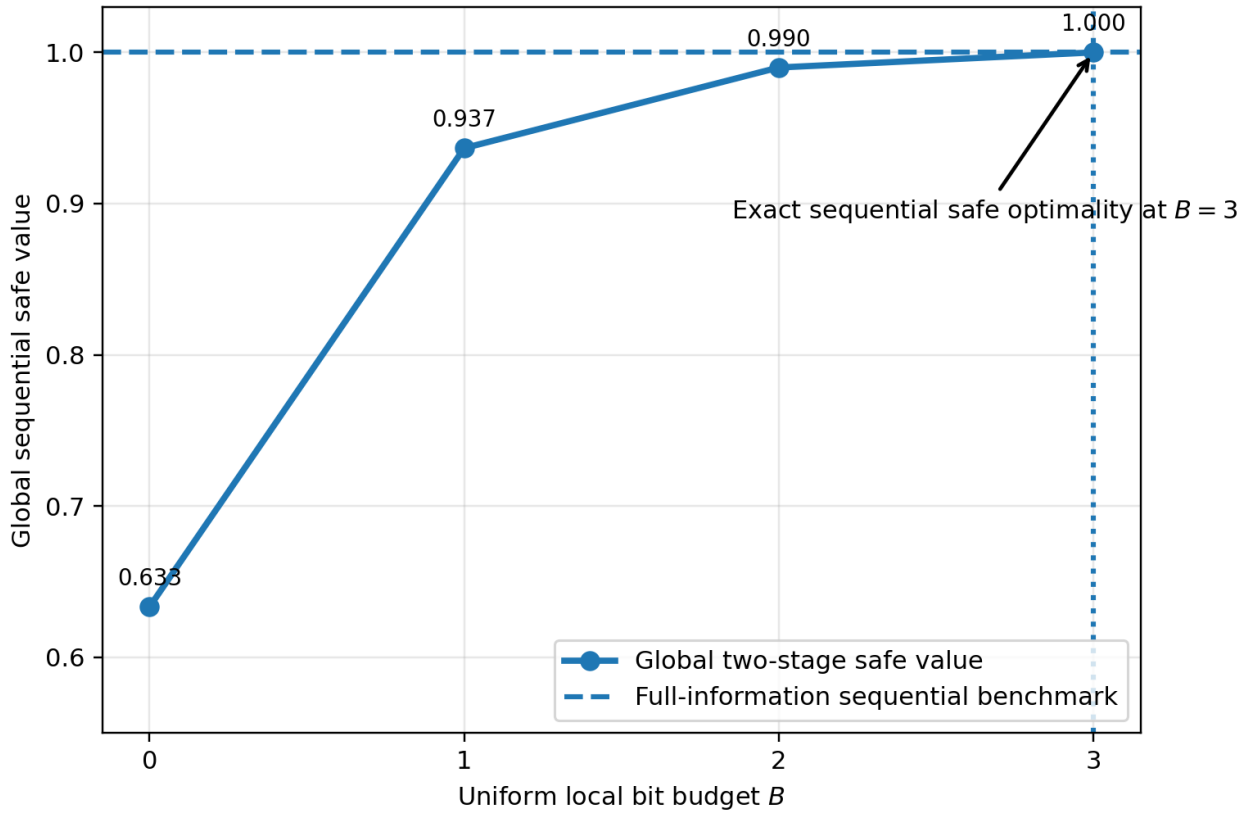


Figure 5. Global two-stage safe value under a uniform local bit budget B applied at all reachable stage-2 histories. Exact full-information sequential performance is attained only when the uniform local budget reaches the largest reachable local lower bound, here $B = 3$.

Table 8. 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

Table 9. 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