STRUCTURALLY CLOSE, SECURITY-WISE FAR: CONTRADICTORY POLICIES IN NEURAL NETWORKS

Anonymous authors
Paper under double-blind review

ABSTRACT

How close can two neural subnetworks be if they implement essentially the same benign functionality but contradictory security policies? We formalize this question through *structural security distance* (SSD), the minimum Hamming distance between masks of a common parent network that approximate a prescribed safe/unsafe policy pair. We also distinguish unrestricted SSD from distance under mask-size constraints and from the distance between a specified safe mask and the nearest unsafe realization, since these quantities need not agree. We establish four main results. First, structural proximity depends on which realizations are compared. We construct a perturbation-stable ReLU family in which the minimum-size safe and unsafe masks are $2q$ edits apart, the minimum safe mask is q edits from the nearest unsafe realization, and the unrestricted SSD is a single mask-coordinate change. Thus, a globally close safe/unsafe pair does not imply that a particular deployed safe model is close to an unsafe realization. Second, low-dimensional safety localization does not imply a small number of structural edits: a one-dimensional safety direction can require $\Theta(d)$ binary neuron-gate changes, and under additional assumptions on benign representations, every discrete one-gate-at-a-time transition can incur a constant benign-error barrier even though continuous directional ablation preserves benign behavior. Third, estimating normalized SSD remains NP-hard in the worst case even at fixed nonzero approximation tolerance, with guaranteed feasible safe and unsafe masks, known minimum sizes, bounded sparse inputs, and robustness to parameter perturbations. Finally, we show that any transition between opposite policies must accumulate sufficient functional change, yielding certificates for the safety of a deployed mask within a bounded number of edits.

1 INTRODUCTION

Modern neural networks can support many functions within one parameterization. Lottery ticket results show that sparse subnetworks can preserve task performance (Frankle & Carbin, 2019). Moreover, strong variants demonstrate that pruning randomly initialized networks can yield useful functions without training the retained weights (Ramanujan et al., 2020; Malach et al., 2020; Pensia et al., 2020). Multi-task supermasks and universal tickets further show that one parent can support many different functions (Wortsman et al., 2020; Burkholz et al., 2021). Security-related sparsity is also well-established: robust tickets, Trojan tickets, and unsafe/safety tickets have all been observed in the last few years (Shi et al., 2022; Zheng et al., 2022; Chen et al., 2022; Si et al., 2026).

These results leave a different question open. Suppose two target policies implement the same functionality but make opposite decisions on a security-sensitive region. Examples include clean versus trigger-conditioned behavior and code-generation policies with identical utility but opposite security constraints. The question then becomes: among all subnetworks of one parent network that realize those policies, how much structure must change?

We study this as a pairwise property of the hypothesis space induced by a fixed parent network. For a fixed parent F_θ, the structural security distance minimizes

$$D^{\mathrm{sec}}_{\varepsilon_s,\varepsilon_u}(f_s, f_u; \theta) = \min_{m_s, m_u} d_{\mathrm{H}}(m_s, m_u) \tag{1}$$

over masks approximating the safe and unsafe targets within tolerances ε_s and ε_u, respectively, where d_H is the Hamming distance. Unlike Hamming or Jaccard similarity between two already-found

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

masks, Eq. (1) optimizes over all feasible masks. Unlike neural network repair, it does not begin from one specific mask. This distinction matters because the closest pair may use non-minimal masks that contain dormant circuitry. We therefore also study structural security distance (SSD) when masks are restricted to remain near their minimum sizes, and the distance from a specified safe mask to the closest unsafe one.

For auditing, the relevant question is whether a specified model can be changed to the opposite policy under the allowed edits and utility constraints. This requires specifying which masks are compared, which edits are allowed, whether benign functionality must be preserved during the edit sequence, and whether the distance is certified or only estimated. We thus distinguish unrestricted SSD, SSD under mask-size constraints, and the distance from a specified safe mask to the closest unsafe realization. Our main contributions are:

1. We show that structural proximity between opposite policies depends on which feasible realizations are compared and on the allowed mask-size budget. We study what happens when each mask may use only a limited number of parameters beyond its minimum size. If every such near-minimum feasible mask contains a minimum-size feasible mask, Theorem 2 gives
$$D^{\mathrm{sec},[b_s,b_u]} \geq \max\{D^{\mathrm{sec},[0]} - b_s - b_u, 0\}. \tag{2}$$
Theorem 3 attains this bound exactly: $D^{\mathrm{sec},[b]} = 2(q-b)$ for $b < q$. At $b = q$, a qualitatively different unsafe realization becomes feasible, and the distance goes from 2 to 1. Overall, the minimum-size distance is $2q$, the unrestricted distance is 1, and the minimum safe mask is q edits from the closest unsafe mask. The separation survives a perturbation radius independent of q.
2. Low-dimensional activation-space control need not imply a small neuron-level structural distance. The condition on the security-domain representations alone already gives a lower bound on the number of binary neuron-gate changes, with the bound determined by how concentrated the safety direction is across coordinates; no assumption on benign representations is needed. Theorem 6 gives an exact family in which ablating the one-dimensional direction $I - vv^\top$ removes the safety behavior while minimum safe and unsafe neuron masks differ in $\Theta(d)$ coordinates. Theorem 7 adds explicit assumptions on benign representations and gives conditions under which every path of individual binary neuron-gate changes incurs a constant benign-error barrier, even when continuous directional ablation preserves benign behavior.
3. Theorem 8 shows that SSD is also difficult to audit in the worst case: at approximation tolerance $1/8$, estimating normalized SSD to additive error $1/8$ is NP-hard even for bounded-input masked linear parents when feasible safe and unsafe masks are guaranteed to exist, and C_s, C_u are known in advance. The same sets of feasible masks remain throughout a fixed ℓ_∞ parameter neighborhood.
4. Theorem 9 lower-bounds the cumulative output change required along any shortest edit path, and Corollary 10 converts uniform one-toggle bounds into a guarantee that a deployed mask remains safe within a specified number of edits.

2 Related Work

Pruning and subnetworks. Lottery ticket and strong ticket results establish the existence of accurate subnetworks inside trained or random parents (Frankle & Carbin, 2019; Ramanujan et al., 2020; Malach et al., 2020; Pensia et al., 2020). Multi-task masks, universal tickets, and mask-overlap methods study coexistence and similarity of subnetworks (Wortsman et al., 2020; Burkholz et al., 2021; You et al., 2025; Sun et al., 2020). Security work has also identified robust, unsafe, and safety-relevant sparse structures; see, e.g., (Shi et al., 2022; Zheng et al., 2022; Chen et al., 2022; Si et al., 2026). Prior pruning theory discusses the computational difficulty of finding optimal subnetworks (Malach et al., 2020). Minimum Hamming distance between feasible solutions has also been studied in combinatorial settings. For example, Behrisch et al. analyze nearest satisfying assignments and pairs of satisfying assignments with minimum Hamming distance in propositional models (Behrisch et al., 2018). Our setting, where feasibility is defined by functional approximation of neural-network policies, has separate safe and unsafe target sets; we also study restrictions on mask size, distance from a specified mask, and security-specific edit constraints.

Activation directions and localized safety mechanisms. Refusal and safety work identifies low-dimensional activation directions where addition or ablation changes behavior (Arditi et al., 2024; Pan et al., 2025), while safety neuron work localizes behavior to specific coordinate components (Chen et al., 2025). Li and Kim's Superficial Safety Alignment Hypothesis further suggests that a small set of safety-critical components can be sufficient to establish safety guardrails in LLMs (Li & Kim, 2026). These results motivate asking whether localization to a small number of components, or to a low-dimensional activation direction, implies that opposite safety policies are separated by only a small number of structural edits. Our results show that these notions need not coincide. Projection-based concept-removal methods operate more generally on linear subspaces. INLP iteratively removes linearly predictive directions (Ravfogel et al., 2020), while RLACE formulates concept erasure as the removal of a learned low-dimensional subspace (Ravfogel et al., 2022). LEACE and related methods provide further closed-form or geometry-aware variants (Belrose et al., 2023; Shi et al., 2026). These approaches act in representation space, while our direction–neuron results ask how such low-dimensional interventions relate to discrete coordinate edits. Provable repair instead starts from one prescribed network and searches for a small change satisfying a specification (Sotoudeh & Thakur, 2021; Fu & Li, 2022; Fu et al., 2024). Recent selective steering methods also aim to isolate more specific behavioral directions. RepIt, for example, constructs concept-specific refusal vectors that allow targeted steering while preserving refusal on other concepts (Siu et al., 2026). Such results further motivate distinguishing the dimensionality of a representation-space intervention from the number of coordinates needed to realize a comparable structural edit.

3 Problem Setup

Let a fixed parent network be $F_\theta : \mathcal{X} \to \mathbb{R}^C$, with parameters $\theta \in \mathbb{R}^P$. A binary mask $m \in \{0,1\}^P$ induces $F_{\theta,m} := F_{m\odot\theta}$, where $\odot$ denotes coordinate-wise multiplication. We use

$$\|m\|_0 := |\{i : m_i = 1\}|, \qquad d_{\mathrm{H}}(m, m') := \|m - m'\|_0, \tag{3}$$

where $\|m\|_0$ is the number of active mask coordinates, d_{H} is their Hamming distance, and $supp(m) := \{i : m_i = 1\}$. When $\mathrm{supp}(m') \subseteq \mathrm{supp}(m)$, we say that m' is a submask of m. Although we write masks over individual parameters for simplicity, one mask coordinate may instead control a predefined group, such as a neuron, attention head, expert, or adapter; then $\|m\|_0$ and d_{H} count active groups and group changes.

For $\mathcal{A} \subseteq \mathcal{X}$, define

$$d_{\mathcal{A}}(f, g) := \sup_{x\in\mathcal{A}} \|f(x) - g(x)\|_\infty. \tag{4}$$

Let $\mathcal{B}$ denote a benign domain and let $\varnothing \neq \mathcal{R} \subseteq \mathcal{X}$ denote a security-sensitive domain.

Definition 1 (Security score and opposite policies)**.** *A security score is a function $s : \mathcal{X} \times \mathbb{R}^C \to \mathbb{R}$ that is L_s-Lipschitz in its second argument. We say that safe target policy f_s and unsafe target policy f_u make opposite security decisions with margin $\gamma > 0$ on $\mathcal{R}$ if*

$$s(x, f_s(x)) \le -\gamma, \qquad s(x, f_u(x)) \ge \gamma \qquad \forall x \in \mathcal{R}. \tag{5}$$

We call the pair benign-equivalent when $f_s = f_u$ on $\mathcal{B}$.

The score can represent a threshold rule, a logit difference, or another task-specific security decision. Unlike a raw output gap, Eq. (5) guarantees opposite decisions. Lipschitzness immediately implies

$$\|f_s(x) - f_u(x)\|_\infty \ge \frac{2\gamma}{L_s} \qquad \forall x \in \mathcal{R}. \tag{6}$$

Moreover, an ε_s-approximation to f_s preserves the safe sign whenever $L_s\varepsilon_s < \gamma$, and analogously for f_u.

For approximation tolerance ε and domain $\mathcal{A}$, let

$$\mathcal{T}_{\varepsilon,\mathcal{A}}(f;\theta) := \{m : d_{\mathcal{A}}(F_{\theta,m}, f) \le \varepsilon\}, \tag{7}$$

and

$$C_{\varepsilon,\mathcal{A}}(f;\theta) := \min_{m\in\mathcal{T}_{\varepsilon,\mathcal{A}}(f;\theta)} \|m\|_0, \tag{8}$$

with $+\infty$ if no feasible mask exists. We abbreviate $\mathcal{T}_\varepsilon := \mathcal{T}_{\varepsilon,\mathcal{X}}$ and $C_\varepsilon := C_{\varepsilon,\mathcal{X}}$. $T_{\varepsilon,A}(f;\theta)$ is the set of subnetworks that realize f within tolerance ε on A, while $C_{\varepsilon,A}(f;\theta)$ is the minimum number of active mask coordinates required by any such realization.

Definition 2 (Structural security distance)**.** *For target policies f_s, f_u and tolerances $\varepsilon_s, \varepsilon_u$,*

$$D^{\mathrm{sec}}_{\varepsilon_s,\varepsilon_u}(f_s, f_u; \theta) := \min_{\substack{m_s \in \mathcal{T}_{\varepsilon_s}(f_s;\theta) \\ m_u \in \mathcal{T}_{\varepsilon_u}(f_u;\theta)}} d_{\mathrm{H}}(m_s, m_u). \tag{9}$$

If either feasible set is empty, the value is $+\infty$.

Here, "distance" means the minimum Hamming distance between the feasible safe and unsafe mask sets; D^{sec} is not intended to define a metric on policies.

Let $C_s = C_{\varepsilon_s}(f_s; \theta)$ and $C_u = C_{\varepsilon_u}(f_u; \theta)$. If both are finite and $C_s + C_u > 0$, define

$$\overline{D}^{\mathrm{sec}}_{\varepsilon_s,\varepsilon_u} := \frac{D^{\mathrm{sec}}_{\varepsilon_s,\varepsilon_u}}{C_s + C_u}; \tag{10}$$

otherwise set $\overline{D}^{\mathrm{sec}} = +\infty$. Since a minimum-size feasible pair has Hamming distance at most $C_s + C_u$, the normalized quantity lies in $[0, 1]$ whenever both feasible sets are nonempty. Using the minimum mask sizes in the denominator prevents a pair from appearing artificially close merely because both masks contain removable common structure. Equivalently, the normalization expresses structural disagreement relative to the minimum combined complexity needed to realize the two policies, rather than as an absolute number of mask-coordinate changes. It does not, however, prevent the masks that minimize D^{sec} from containing dormant circuitry.

To restrict attention to masks near their minimum sizes, for integer-valued budgets $b_s, b_u \geq 0$ define

$$D^{\mathrm{sec},[b_s,b_u]}_{\varepsilon_s,\varepsilon_u} := \min_{\substack{m_s \in \mathcal{T}_{\varepsilon_s}(f_s),\ \|m_s\|_0 \leq C_s + b_s \\ m_u \in \mathcal{T}_{\varepsilon_u}(f_u),\ \|m_u\|_0 \leq C_u + b_u}} d_{\mathrm{H}}(m_s, m_u). \tag{11}$$

For equal budgets write $D^{\mathrm{sec},[b]} := D^{\mathrm{sec},[b,b]}$ and $D^{\mathrm{sec},[\infty]} = D^{\mathrm{sec}}$, and set

$$\overline{D}^{\mathrm{sec},[b]} := \frac{D^{\mathrm{sec},[b]}}{C_s + C_u}. \tag{12}$$

For a specified feasible safe mask $m_s^\star$, we define its distance to the closest unsafe mask as

$$D^{\mathrm{fixed}}(m_s^\star) := \min_{m_u \in \mathcal{T}_{\varepsilon_u}(f_u)} d_{\mathrm{H}}(m_s^\star, m_u). \tag{13}$$

A general stability lemma in Appendix A.2 shows that if every mask lies a positive discrepancy margin away from the feasibility threshold, any uniformly smaller parameter perturbation preserves the entire set of feasible masks.

4 Size Budgets and Distance from a Fixed Mask

Proposition 1 (Minimum-size opposite policies can be one edge apart)**.** *For every $n \geq 1$ and $\gamma > 0$, there are benign-equivalent targets that make opposite security decisions with margin γ and have unique exact masks of sizes n and $n + 1$ such that*

$$D^{\mathrm{sec}}_{0,0} = D^{\mathrm{sec},[0]}_{0,0} = 1, \qquad \overline{D}^{\mathrm{sec}}_{0,0} = \frac{1}{2n+1}. \tag{14}$$

A masked linear construction and proof are in Appendix A.3. This example motivates asking whether a small optimized distance persists when masks must remain near minimum size, or when the safe mask is fixed in advance.

4.1 Distance Under a Size Budget

The unrestricted optimum can answer a different question from the one relevant to a compressed or deployed model: a nonminimal mask may already contain almost all circuitry needed for the opposite policy. The theorem below bounds how much the distance can decrease under this near-minimum-size condition.

Theorem 2 (Effect of allowing extra active parameters)**.** *Let $D_0 := D^{\mathrm{sec},[0,0]}_{\varepsilon_s,\varepsilon_u}(f_s, f_u; \theta)$. Suppose every safe feasible mask m_s with $\|m_s\|_0 \le C_s + b_s$ contains a minimum-size feasible safe submask, and every unsafe feasible mask m_u with $\|m_u\|_0 \le C_u + b_u$ contains a minimum-size feasible unsafe submask. Then*

$$D^{\mathrm{sec},[b_s,b_u]}_{\varepsilon_s,\varepsilon_u} \ge \max\{D_0 - b_s - b_u, 0\}. \tag{15}$$

In particular, if the same statement holds with budget b for both safe and unsafe masks,

$$D^{\mathrm{sec},[b]} \ge \max\{D^{\mathrm{sec},[0]} - 2b, 0\}. \tag{16}$$

Thus, adding one extra active parameter to either mask can reduce their Hamming distance by at most one as long as no qualitatively different feasible realization becomes available.

The proof is a Hamming-triangle argument (Appendix A.1). The following family attains the bound exactly until a different unsafe realization becomes feasible.

Theorem 3 (Exact distance–budget tradeoff)**.** *Fix $0 < \gamma \le 1/2$ and $0 < \varepsilon < \gamma/4$. For every integer $q \ge 3$, there exists a ReLU parent with one hidden layer, four hidden neurons, two outputs, and exactly $4q$ independently maskable scalar parameters, together with benign-equivalent targets (f_s, f_u), such that:*

1. *all parent weights have magnitude at most one; input and target-output norms are bounded independently of q;*
2. *the common benign functionality is nonconstant, and on $\mathcal{R}$ the targets make opposite sign decisions with margin at least $\gamma/2$;*
3. *$C_s = C_u = 2q$; for every $0 \le b < q$, each feasible safe or unsafe mask of size at most $2q + b$ contains a minimum-size feasible mask of the same type, while this property fails for the unsafe masks at $b = q$; and for every integer $b \ge 0$,*

$$D^{\mathrm{sec},[b]}_{\varepsilon,\varepsilon} = \begin{cases} 2(q-b), & 0 \le b < q, \\ 1, & b \ge q; \end{cases} \tag{17}$$

4. *the minimum-size safe mask $m_s^\star$ is unique and*

$$D^{\mathrm{fixed}}(m_s^\star) = q. \tag{18}$$

Consequently,

$$\overline{D}^{\mathrm{sec},[\infty]} = \frac{1}{4q} \to 0, \qquad \overline{D}^{\mathrm{sec},[0]} = \frac{1}{2}, \qquad \frac{D^{\mathrm{fixed}}(m_s^\star)}{C_s + C_u} = \frac{1}{4}. \tag{19}$$

For every fixed relative overhead $0 \le \alpha < 1/2$,

$$\overline{D}^{\mathrm{sec},[\lfloor 2\alpha q \rfloor]} \ge \frac{1}{2} - \alpha > 0. \tag{20}$$

The construction uses one q-parameter utility branch U and three q-parameter security branches A, B, C. The safe target requires $U \cup A$; the unsafe target can use either $U \cup B$ or $U \cup A \cup C$. With fewer than q additional parameters, the latter unsafe realization is unavailable, and the A/B branches remain separated. Once q additional parameters are allowed, a safe mask may contain an almost-complete dormant C branch, after which enabling one output edge switches policy. Appendix A.4 characterizes all feasible masks and provides the exact formula.

Corollary 4 (Perturbation-stable separation)**.** *Applying Proposition 11 to the same construction, perturb every existing parent parameter by at most*

$$\|\widetilde{\theta} - \theta\|_\infty \le \varepsilon/42. \tag{21}$$

Then the feasible mask sets remain exactly unchanged:

$$\mathcal{T}_\varepsilon(f_j; \widetilde{\theta}) = \mathcal{T}_0(f_j; \theta), \qquad j \in \{s, u\}. \tag{22}$$

Hence every distance formula in Theorem 3 survives a positive perturbation radius independent of q, and the feasible safe and unsafe masks still make opposite security decisions with margin greater than $\gamma/4$.

Note that Proposition 1 and Theorem 3 are complementary. A one-edge distance can be genuine even at minimum size, but unrestricted one-edge distance can also happen because a nonminimal mask carries dormant capacity. Normalizing by the minimum mask sizes removes denominator padding, but it does not make the unrestricted optimum representative of compressed models.

A small unrestricted D^{sec} certifies only that some safe/unsafe pair is close. It does not certify a small change from a deployed mask. A deployment-oriented audit should therefore report the distance from the specified mask and the allowed mask-size budget, rather than only the globally closest pair.

5 Directional Control Versus Binary Neuron Edits

Low-dimensional activation-space control need not imply small neuron-level structural distance. Proposition 5 gives a lower bound using only the security-domain representations and makes no assumption on benign representations. Theorem 6 gives an exact construction, while Theorem 7 adds assumptions on benign representations to obtain a benign-error barrier along discrete edit paths. Full proofs are in the appendix.

Proposition 5 (Gate-distance bound from security-domain behavior)**.** *Let*

$$F_m(x) = g(x) + (m \odot v)^\top h(x), \qquad m \in \{0,1\}^d, \quad \|v\|_2 = 1, \tag{23}$$

with $f_s = F_{\mathbf{1}}$, $f_u = F_{\mathbf{0}} = g$, and nonempty $\mathcal{R}$. Assume

$$\sup_{x \in \mathcal{R}} \|h(x) - v\|_2 \le \rho. \tag{24}$$

For approximation tolerance $\varepsilon \ge 0$, define

$$Q_{\varepsilon,\rho} := \left(\frac{\rho + \sqrt{\rho^2 + 4\varepsilon}}{2} \right)^2, \qquad \mu(v) := d\|v\|_\infty^2. \tag{25}$$

Here, $\mu(v)$ measures coordinate concentration: $\mu(v) \approx 1$ when the direction is distributed nearly uniformly across coordinates, with larger values indicating concentration on fewer coordinates. Then every ε-feasible safe/unsafe pair in this gate family satisfies

$$D_{\varepsilon,\varepsilon}^{\mathrm{sec,gate}} \ge \left\lceil \frac{d}{\mu(v)} \max\{1 - 2Q_{\varepsilon,\rho}, 0\} \right\rceil. \tag{26}$$

No spanning or coverage assumption on the benign representations is required.
Let us consider the following example. For uniform v, $d = 1000$, $\varepsilon = 0.1$, and $\rho = 0.05$, the bound gives at least 766 binary neuron-gate edits.

Thus, no assumption on benign representations is needed for this distance lower bound. Such assumptions are used only for the stronger benign-error statement below.

Theorem 6 (Exact direction–neuron separation)**.** *For every $d \ge 2$ there is a fixed d-neuron gated linear representation, with identical nonconstant benign functionality and opposite security decisions, such that for every $0 \le \varepsilon < 1/2$, with*

$$r_\varepsilon = \left\lfloor \frac{d}{2} \left(1 - \sqrt{1 - 4\varepsilon^2} \right) \right\rfloor, \tag{27}$$

its neuron-gated distances satisfy

$$D_{\varepsilon,\varepsilon}^{\mathrm{sec}} = d - 2r_\varepsilon, \qquad D_\varepsilon^{\mathrm{fixed}} = d - r_\varepsilon. \tag{28}$$

Ablating the one-dimensional direction v via the orthogonal projection $I - vv^\top$ onto $v^\perp$ switches the policy while preserving every benign output. Every path from the exact safe mask to the exact unsafe mask that changes one neuron gate per step incurs

$$\max_{m \in \pi} d_{\mathcal{B}}(F_m, f_s) \ge \frac{\sqrt{\lfloor d^2/4 \rfloor}}{d}, \tag{29}$$

and this bound is tight, whereas the continuous directional path has zero benign error. At zero tolerance, $\overline{D}^{\mathrm{sec}} = d/(d+2) \to 1$. For an arbitrary unit safety direction v, the same gate family obeys

$$D_{\varepsilon,\varepsilon}^{\mathrm{sec}} \ge \left\lceil \frac{d}{\mu(v)} \sqrt{1 - 4\varepsilon^2} \right\rceil, \qquad \mu(v) := d\|v\|_\infty^2. \tag{30}$$

A safety mechanism can be one-dimensional in activation space yet require changes to a linear fraction of neuron gates when edits are restricted to binary coordinates. Therefore, directional dimension alone does not control neuron-level distance: coordinate concentration does. A distributed direction with $\mu(v) \approx 1$ can be one-dimensional in activation space while requiring $\Theta(d)$ neuron edits. Theorem 6 treats a one-dimensional direction, but the same distinction motivates a broader question for multi-direction methods such as INLP or RLACE: a low-dimensional subspace may still be distributed across many coordinates. Extending the structural bound from a single direction to a k-dimensional subspace would require controlling the coordinate concentration of the subspace, rather than its dimension alone.

Theorem 7 (Direction–neuron separation with benign constraints)**.** *Let*

$$F_m(x) = g(x) + (m \odot v)^\top h(x), \qquad m \in \{0,1\}^d, \quad \|v\|_2 = 1, \tag{31}$$

with $f_s = F_{\mathbf{1}}$ and $f_u = F_{\mathbf{0}} = g$. Assume $\mathcal{R} \neq \varnothing$ and, for constants $\beta \geq 0$, $a > 0$, and $\rho \geq 0$,

$$\sup_{\mathcal{B}} |v^\top h| \leq \beta, \qquad \sup_{\mathcal{B}} |w^\top h| \geq a\|w\|_2 \; \forall w \perp v, \qquad \sup_{\mathcal{R}} \|h - v\|_2 \leq \rho. \tag{32}$$

Set $\nu = \|v\|_\infty^2$, $\mu = d\nu$, and $\eta = (\varepsilon + \beta/2)/a$. If $\varepsilon + \rho < 1/2$ and $\eta < 1/2$, then the structural distance restricted to these gates satisfies

$$D_{\varepsilon,\varepsilon}^{\mathrm{sec,gate}} \geq \left\lceil \frac{d}{\mu}\sqrt{1 - 4\eta^2} \right\rceil. \tag{33}$$

The directional-ablation path $h_t = (I - tvv^\top)h$ reaches the unsafe target with benign error at most β throughout, while every path that changes one binary gate per step between feasible safe and unsafe masks contains a mask with

$$d_{\mathcal{B}}(F_m, f_s) \geq \left\lceil \frac{a\sqrt{1-\nu^2} - \beta(1+\nu)}{2} \right\rceil_+. \tag{34}$$

For uniform v and even d, the sharper barrier is $(a - \beta)/2$. Moreover, $d_{\mathcal{B}}(f_s, f_u) \leq \beta$, and the feasible safe and unsafe masks retain a margin of at least $1/2 - \rho - \varepsilon$ under the natural threshold score.

Hence, continuous directional editing can preserve benign behavior, while every discrete one-gate-at-a-time route between the same policies must pass through a model with nontrivial benign error.

The barrier concerns binary one-gate edits: continuous scaling $m(t) = (1-t)\mathbf{1}$ traces the same output path as directional ablation. For uniform v and even d, the edit classes are separated by any benign budget $\beta \leq \zeta < (a - \beta)/2$, which exists when $a > 3\beta$. Coverage requires the projected benign representations to quantitatively span $v^\perp$; a finite benign set with fewer than $d - 1$ linearly independent projected vectors cannot have $a > 0$. The theorem concerns the specified gated layer, since other editable circuitry may provide alternative routes. Projection/directional ablation itself is established (Arditi et al., 2024; Belrose et al., 2023; Shi et al., 2026); our contribution is the quantitative structural and path separation.

Directional dimension and binary coordinate-edit cost describe different kinds of edits. An audit that reports a safety direction should therefore state which edits are allowed. Proposition 5 uses only the condition on the security domain; Theorem 7 needs the additional assumptions on benign representations to obtain a benign-error barrier.

6 Hardness of Approximation

We next study the worst-case complexity of estimating structural security distance.

Theorem 8 (Hardness at fixed approximation tolerance)**.** *At approximation tolerance $\varepsilon_0 = 1/8$, it is NP-hard to estimate $\overline{D}^{\mathrm{sec}}_{\varepsilon_0,\varepsilon_0}$ to additive error $1/8$ for a family of explicitly represented finite rational instances with unit-weight one-output masked linear parents whose evaluation inputs have at most four $\{\pm 1\}$ entries. In every such instance, the targets are benign-equivalent and make opposite security decisions with margin $1/2$ on a nonempty security domain; both feasible mask sets are guaranteed nonempty, and C_s, C_u are known in advance. Moreover, for each instance produced by*

the reduction, every parent-weight perturbation satisfying $\|\Delta\theta\|_\infty \le 1/64$ preserves both feasible mask sets at tolerance $1/8$; hence the unrestricted, minimum-size, and fixed-safe-mask distances are unchanged throughout this parameter ball. The same additive-hardness applies after normalization by $C_s + C_u$.

The constant $1/8$ is chosen for convenience: the reduction has integer-valued feasibility constraints with unit violation gap, while the normalized YES/NO SSD values differ by more than $1/4$, making additive error $1/8$ preserve both feasibility and distinguishability. The proof in Appendix D reduces from the NP-complete Restricted Exact Cover by 3-Sets variant in which every element occurs in exactly three sets (Gonzalez, 1985). RX3C asks whether a collection of 3-element sets contains an exact cover of a universe of $3q$ elements, under the restriction that every universe element occurs in exactly three sets. The reduction uses integral inputs with at most four nonzero entries, which gives both a normalized YES/NO gap greater than $1/4$ and the fixed perturbation radius.

Guaranteed feasibility and known minimum mask sizes do not make SSD easy to estimate in the worst case. Audits should therefore distinguish certified bounds from heuristic estimates of D^{sec}.

7 Influence Along Edit Paths

For a mask m and coordinate e, write $m^{\oplus e}$ for the mask obtained by flipping coordinate e, and define the resulting output change on the security domain by

$$\lambda_{\mathcal{R}}(m, e) := d_{\mathcal{R}}(F_{\theta,m}, F_{\theta,m^{\oplus e}}). \tag{35}$$

Let $\mathcal{P}_{\mathrm{sp}}(m, m')$ denote the shortest Hamming paths from m to m'. For a path $\pi = (m^{(0)}, e_1, m^{(1)}, \ldots, e_k, m^{(k)})$, define its cumulative output change

$$L_{\mathcal{R}}(\pi) = \sum_{j=1}^{k} \lambda_{\mathcal{R}}(m^{(j-1)}, e_j). \tag{36}$$

Theorem 9 (Required change along a shortest edit path)**.** *Suppose (f_s, f_u) satisfy Eq. (5) for an L_s-Lipschitz security score. For feasible masks $m_s \in \mathcal{T}_{\varepsilon_s}(f_s)$ and $m_u \in \mathcal{T}_{\varepsilon_u}(f_u)$, define*

$$\tau := \max\left\{\frac{2\gamma}{L_s} - \varepsilon_s - \varepsilon_u, 0\right\}. \tag{37}$$

Then every shortest Hamming path $\pi \in \mathcal{P}_{\mathrm{sp}}(m_s, m_u)$ satisfies

$$L_{\mathcal{R}}(\pi) \ge \tau. \tag{38}$$

If $k = d_{\mathrm{H}}(m_s, m_u) \ge 1$ and $\tau > 0$, every such path contains at least one toggle with

$$\lambda_{\mathcal{R}}(m^{(j-1)}, e_j) \ge \frac{\tau}{k}. \tag{39}$$

The cumulative lower bound holds for any nonempty class of paths. The per-toggle bound τ/k additionally requires a nonempty subclass of shortest Hamming paths; for a path of length ℓ, the corresponding bound is τ/ℓ.

Theorem 9 is a path-wise statement. It does not identify an influential coordinate at either endpoint, require the same coordinate to be important on different paths, or guarantee benign behavior of intermediate masks.

Corollary 10 (Guarantee for a deployed mask)**.** *Suppose a deployed mask m_0 satisfies*

$$s(x, F_{\theta,m_0}(x)) \le -\gamma \qquad \forall x \in \mathcal{R}. \tag{40}$$

For each coordinate e, let $c_e \ge 0$ satisfy

$$d_{\mathcal{R}}(F_{\theta,m}, F_{\theta,m^{\oplus e}}) \le c_e \tag{41}$$

for every intermediate mask m within Hamming distance at most $k-1$ of m_0. Let $c_{(1)} \ge \cdots \ge c_{(P)}$ be these bounds sorted in descending order. If

$$L_s \sum_{i=1}^{k} c_{(i)} < \gamma, \tag{42}$$

then every mask differing from m_0 in at most k coordinates preserves the safe sign on all of $\mathcal{R}$.

Indeed, a shortest edit path changes at most k distinct coordinates, so telescoping bounds the output change by the sum of the corresponding c_e, which is at most the sum of the k largest bounds; Lipschitzness then leaves the score strictly below zero. Gradients or saliency values computed only at m_0 do not by themselves provide the required bounds for intermediate masks.

Appendix A.7 gives a related consequence for the rescaling-invariant ReLU path metric of Gonon et al. (2025): an independently bounded one-toggle change yields a lower bound on the number of mask changes required.

8 Supporting Upper Bounds from Shared Computation

We also give conditions under which close safe and unsafe masks exist. Appendix B shows that if $f_s = g + h_s$ and $f_u = g + h_u$ share benign computation g and the policy-specific residuals have sparse tickets of sizes K_s, K_u, then confidence amplification gives, with high probability, $D^{\mathrm{sec}} \le K_s + K_u$ and $\overline{D}^{\mathrm{sec}} \le (K_s + K_u)/(2A_*)$, where $A_* > 0$ lower-bounds the size needed to approximate g. The same appendix gives a fixed-tolerance two-layer random-ReLU family with $D^{\mathrm{sec}} = 1$ and $\overline{D}^{\mathrm{sec}} = O(\log W/W)$ while retaining a nonvanishing security margin. These are existence results, not a claim that increasing width makes a fixed deployed model less safe.

9 Discussion

Proposition 5 and Theorems 6 and 7 show why neuron-level structural distance and activation-space localization are not interchangeable. The condition on the security-domain representations, together with the coordinate concentration of the safety direction, already lower-bounds the number of binary gate changes, even for rank-deficient or finite benign evaluation sets. Additional assumptions on benign representations are needed only for the stronger claim that discrete edit paths incur benign error. Because mask coordinates can represent groups, the same distinction applies to heads, adapters, and expert blocks. Empirically identified safety neurons (Chen et al., 2025) can be studied by restricting which gates may change, while direction-level analyses (Arditi et al., 2024; Pan et al., 2025) describe geometry in activation space. NeuroStrike (Wu et al., 2026) shows that changing a small set of safety neurons can substantially affect aligned refusal behavior. Our results address the complementary theoretical question of when such localization implies a small structural distance.

Limitations. The main results concern representational possibilities rather than training dynamics or realistic attack accessibility. Proposition 1 and Theorem 3 are explicit constructions; Theorem 9 gives necessary rather than sufficient conditions. The direction results fix a representation and a specified gated layer, so other editable circuitry can provide alternative routes. Proposition 5 requires no assumption on benign representations, while the stronger benign-error result in Theorem 7 does. The random-parent upper bounds require sparse-ticket assumptions and do not imply that increasing width makes a fixed task less safe. The hardness theorem is a worst-case result and does not rule out efficient estimation for structured families of trained models. It shows that additive estimation at fixed tolerance can remain NP-hard even with bounded inputs, guaranteed feasible safe and unsafe masks, known minimum mask sizes, and a fixed-radius parameter neighborhood.

10 Conclusion

The interpretation of structural security distance depends on the audit setting. A close safe/unsafe pair does not imply that a specified deployed mask is close to an unsafe realization, and a low-dimensional safety direction does not imply a small number of neuron-gate changes. In the worst case, normalized SSD is hard to approximate, while uniform one-toggle bounds can still certify safety within a specified number of edits. An SSD audit should therefore state which masks are compared, which edits are allowed, what utility constraints apply, and whether the reported distance is certified or estimated.

Reproducibility Statement

This paper is theoretical and uses no experimental datasets or code. We state assumptions explicitly and provide complete proofs of all main results and auxiliary extensions in the appendix.

Ethics Statement

The paper develops abstract theory about structural proximity between neural-network policies. It does not provide exploit code, attack datasets, or procedures for compromising deployed systems. The constructions are intended to clarify what structural distance can and cannot certify and may support defensive auditing of concentrated security mechanisms.

AI Use Statement

Generative AI tools were used to help refine the conceptual framework, revise and check mathematical claims, identify and summarize related literature, clarify limitations, and edit the manuscript. No experiments or synthetic datasets were generated for this work. The AI-assisted mathematical statements and proofs were reviewed and checked by the authors, and literature claims were checked against cited sources. The authors take responsibility for the final content of the submission, including all elements produced with the aid of generative AI.

References

Andy Arditi, Oscar Obeso, Aaquib Syed, Daniel Paleka, Nina Panickssery, Wes Gurnee, and Neel Nanda. Refusal in language models is mediated by a single direction. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-4322. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/f545448535dfde4f9786555403ab7c49-Abstract-Conference.html.

Mike Behrisch, Miki Hermann, Stefan Mengel, and Gernot Salzer. Minimal distance of propositional models. *Theory of Computing Systems*, 63(6):1131–1184, December 2018. ISSN 1433-0490. doi: 10.1007/s00224-018-9896-8. URL http://dx.doi.org/10.1007/s00224-018-9896-8.

Nora Belrose, David Schneider-Joseph, Shauli Ravfogel, Ryan Cotterell, Edward Raff, and Stella Biderman. LEACE: Perfect linear concept erasure in closed form. In *Advances in Neural Information Processing Systems*, volume 36, pp. 66044–66063, 2023. doi: 10.52202/075280-2884. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/d066d21c619d0a78c5b557fa3291a8f4-Abstract-Conference.html.

Rebekka Burkholz, Nilanjana Laha, Rajarshi Mukherjee, and Alkis Gotovos. On the existence of universal lottery tickets. *CoRR*, abs/2111.11146, 2021. URL https://arxiv.org/abs/2111.11146.

Jianhui Chen, Xiaozhi Wang, Zijun Yao, Yushi Bai, Lei Hou, and Juanzi Li. Towards understanding safety alignment: A mechanistic perspective from safety neurons. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=AAXMcAyNF6.

Tianlong Chen, Zhenyu Zhang, Yihua Zhang, Shiyu Chang, Sijia Liu, and Zhangyang Wang. Quarantine: Sparsity can uncover the trojan attack trigger for free, 2022. URL https://arxiv.org/abs/2205.11819.

Jonathan Frankle and Michael Carbin. The lottery ticket hypothesis: Finding sparse, trainable neural networks. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=rJl-b3RcF7.

Feisi Fu and Wenchao Li. Sound and complete neural network repair with minimality and locality guarantees. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=xS8AMYiEav3.

Feisi Fu, Zhilu Wang, Weichao Zhou, Yixuan Wang, Jiameng Fan, Chao Huang, Qi Zhu, Xin Chen, and Wenchao Li. Reglo: Provable neural network repair for global robustness properties. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(11):12061–12071, March 2024. ISSN 2159-5399. doi: 10.1609/aaai.v38i11.29094. URL http://dx.doi.org/10.1609/aaai.v38i11.29094.

Antoine Gonon, Nicolas Brisebarre, Elisa Riccietti, and Rémi Gribonval. A rescaling-invariant Lipschitz bound based on path-metrics for modern ReLU network parameterizations. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 20047–20074. PMLR, 13–19 Jul 2025. URL https://proceedings.mlr.press/v267/gonon25a.html.

Teofilo F. Gonzalez. Clustering to minimize the maximum intercluster distance. *Theoretical Computer Science*, 38:293–306, 1985. ISSN 0304-3975. doi: https://doi.org/10.1016/0304-3975(85)90224-5. URL https://www.sciencedirect.com/science/article/pii/0304397585902245.

Jianwei Li and Jung-Eun Kim. Superficial safety alignment hypothesis, 2026. URL https://arxiv.org/abs/2410.10862.

Eran Malach, Gilad Yehudai, Shai Shalev-Schwartz, and Ohad Shamir. Proving the lottery ticket hypothesis: Pruning is all you need. In Hal Daumé III and Aarti Singh (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 6682–6691. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/malach20a.html.

Emanuele Natale, Davide Ferré, Giordano Giambartolomei, Frédéric Giroire, and Frederik Mallmann-Trenn. On the sparsity of the strong lottery ticket hypothesis. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 40565–40592. Curran Associates, Inc., 2024. doi: 10.52202/079017-1283. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/47908cab4e5b696d7af5c7de69f3b7d2-Paper-Conference.pdf.

Wenbo Pan, Zhichao Liu, Qiguang Chen, Xiangyang Zhou, Yu Haining, and Xiaohua Jia. The hidden dimensions of LLM alignment: A multi-dimensional analysis of orthogonal safety directions. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 47697–47716. PMLR, 13–19 Jul 2025. URL https://proceedings.mlr.press/v267/pan25f.html.

Ankit Pensia, Shashank Rajput, Alliot Nagle, Harit Vishwakarma, and Dimitris Papailiopoulos. Optimal lottery tickets via subset sum: Logarithmic over-parameterization is sufficient. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 2599–2610. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/1b742ae215adf18b75449c6e272fd92d-Paper.pdf.

Vivek Ramanujan, Mitchell Wortsman, Aniruddha Kembhavi, Ali Farhadi, and Mohammad Rastegari. What's hidden in a randomly weighted neural network?, 2020. URL https://arxiv.org/abs/1911.13299.

Shauli Ravfogel, Yanai Elazar, Hila Gonen, Michael Twiton, and Yoav Goldberg. Null it out: Guarding protected attributes by iterative nullspace projection. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 7237–7256, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.647. URL https://aclanthology.org/2020.acl-main.647/.

Shauli Ravfogel, Michael Twiton, Yoav Goldberg, and Ryan D Cotterell. Linear adversarial concept erasure. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), *Proceedings of the 39th International Conference on Machine Learning*,

volume 162 of *Proceedings of Machine Learning Research*, pp. 18400–18421. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/ravfogel22a.html.

Chuancheng Shi, Wenhua Wu, Fei Shen, Xiaogang Zhu, Kun Hu, and Zhiyong Wang. Orthoeraser: Coupled-neuron orthogonal projection for concept erasure, 2026. URL https://arxiv.org/abs/2603.11493.

Xupeng Shi, Pengfei Zheng, A. Adam Ding, Yuan Gao, and Weizhong Zhang. Finding dynamics preserving adversarial winning tickets. In Gustau Camps-Valls, Francisco J. R. Ruiz, and Isabel Valera (eds.), *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pp. 510–528. PMLR, 28–30 Mar 2022. URL https://proceedings.mlr.press/v151/shi22a.html.

Wai Man Si, Mingjie Li, Michael Backes, and Yang Zhang. Pruning unsafe tickets: A resource-efficient framework for safer and more robust LLMs. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 26285–26302, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-390-6. doi: 10.18653/v1/2026.acl-long.1209. URL https://aclanthology.org/2026.acl-long.1209/.

Vincent Siu, Nathan W. Henry, Nicholas Crispino, Yang Liu, Dawn Song, and Chenguang Wang. Repit: Steering language models with concept-specific refusal vectors. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=fsZkx8gek0.

Matthew Sotoudeh and Aditya V. Thakur. Provable repair of deep neural networks, 2021. URL https://arxiv.org/abs/2104.04413.

Tianxiang Sun, Yunfan Shao, Xiaonan Li, Pengfei Liu, Hang Yan, Xipeng Qiu, and Xuanjing Huang. Learning sparse sharing architectures for multiple tasks. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):8936–8943, April 2020. ISSN 2159-5399. doi: 10.1609/aaai.v34i05.6424. URL http://dx.doi.org/10.1609/aaai.v34i05.6424.

Mitchell Wortsman, Vivek Ramanujan, Rosanne Liu, Aniruddha Kembhavi, Mohammad Rastegari, Jason Yosinski, and Ali Farhadi. Supermasks in superposition. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 15173–15184. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/ad1f8bb9b51f023cdc80cf94bb615aa9-Paper.pdf.

Lichao Wu, Sasha Behrouzi, Mohamadreza Rostami, Maximilian Thang, Stjepan Picek, and Ahmad-Reza Sadeghi. Neurostrike: Neuron-level attacks on aligned llms. In *Proceedings 2026 Network and Distributed System Security Symposium*, NDSS 2026. Internet Society, 2026. doi: 10.14722/ndss.2026.230660. URL http://dx.doi.org/10.14722/ndss.2026.230660.

Haoran You, Chaojian Li, Pengfei Xu, Yonggan Fu, Yue Wang, Xiaohan Chen, Richard G. Baraniuk, Zhangyang Wang, and Yingyan Celine Lin. Drawing early-bird tickets: Towards more efficient training of deep networks, 2025. URL https://arxiv.org/abs/1909.11957.

Rui Zheng, Bao Rong, Yuhao Zhou, Di Liang, Sirui Wang, Wei Wu, Tao Gui, Qi Zhang, and Xuanjing Huang. Robust lottery tickets for pre-trained language models. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (eds.), *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2211–2224, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.157. URL https://aclanthology.org/2022.acl-long.157/.

A Proofs for the Main Results

A.1 Proof of Theorem 2

Proof. Take arbitrary masks m_s, m_u satisfying the size constraints in Eq. (11). By the theorem assumptions, choose minimum-size feasible submasks $m_s^0 \subseteq m_s$ and $m_u^0 \subseteq m_u$. Then

$$d_{\mathrm{H}}(m_s, m_s^0) = \|m_s\|_0 - C_s \le b_s, \qquad d_{\mathrm{H}}(m_u, m_u^0) = \|m_u\|_0 - C_u \le b_u. \tag{43}$$

The Hamming triangle inequality gives

$$\begin{aligned} D_0 &\le d_{\mathrm{H}}(m_s^0, m_u^0) && (44)\\ &\le d_{\mathrm{H}}(m_s^0, m_s) + d_{\mathrm{H}}(m_s, m_u) + d_{\mathrm{H}}(m_u, m_u^0) && (45)\\ &\le b_s + d_{\mathrm{H}}(m_s, m_u) + b_u. && (46) \end{aligned}$$

Thus $d_{\mathrm{H}}(m_s, m_u) \ge D_0 - b_s - b_u$ for every pair satisfying the size constraints. Taking the minimum and truncating at zero proves Eq. (15); the symmetric form is immediate. □

A.2 Feasible-set stability under parameter perturbations

Proposition 11 (Feasible-set stability)**.** *For a target f, define*

$$\mu_\varepsilon(f;\theta) := \min_{m\in\{0,1\}^P} \left| d_{\mathcal{X}}(F_{\theta,m}, f) - \varepsilon \right|. \tag{47}$$

If $\mu_\varepsilon(f;\theta) > 0$ and

$$\sup_m d_{\mathcal{X}}(F_{\widetilde{\theta},m}, F_{\theta,m}) < \mu_\varepsilon(f;\theta), \tag{48}$$

then $\mathcal{T}_\varepsilon(f;\widetilde{\theta}) = \mathcal{T}_\varepsilon(f;\theta)$.

Proof. For every mask m, triangle inequality implies

$$\left| d_{\mathcal{X}}(F_{\widetilde{\theta},m}, f) - d_{\mathcal{X}}(F_{\theta,m}, f) \right| \le d_{\mathcal{X}}(F_{\widetilde{\theta},m}, F_{\theta,m}) < \mu_\varepsilon. \tag{49}$$

Because the finite mask set has no discrepancy value within μ_ε of the threshold ε, every originally feasible mask has discrepancy at most $\varepsilon - \mu_\varepsilon$ and remains feasible, while every originally infeasible mask has discrepancy at least $\varepsilon + \mu_\varepsilon$ and remains infeasible. Thus, the feasible mask set is unchanged. □

A.3 Proof of Proposition 1

Proof. We give the construction explicitly. Let

$$X = [0,1]^n \times \{0,1\}, \qquad B = [0,1]^n \times \{0\}, \qquad R = [0,1]^n \times \{1\}.$$

For a binary mask

$$m = (m_1, \dots, m_{n+1}) \in \{0,1\}^{n+1},$$

define the masked linear parent

$$F_{\theta,m}(x,t) = \sum_{i=1}^{n} m_i x_i + 2\gamma m_{n+1} t,$$

where the underlying parameter vector is

$$\theta = (1, \dots, 1, 2\gamma).$$

Define the two target policies by

$$f_s(x,t) = \sum_{i=1}^{n} x_i, \qquad f_u(x,t) = \sum_{i=1}^{n} x_i + 2\gamma t,$$

and define the security score by

$$s((x,t),y) = y - \sum_{i=1}^{n} x_i - \gamma.$$

The score is 1-Lipschitz in its prediction argument y. On the benign region B,

$$f_s(x,0) = f_u(x,0) = \sum_{i=1}^{n} x_i.$$

Thus the two targets are benign-equivalent.

On the risk region R,

$$s((x,1), f_s(x,1)) = -\gamma, \qquad s((x,1), f_u(x,1)) = +\gamma.$$

Hence the two targets make opposite security decisions with margin γ.

It remains to characterize the exact masks. Let e_i denote the i-th standard basis vector in $\mathbb{R}^n$. At the benign point $(e_i, 0)$,

$$F_{\theta,m}(e_i, 0) = m_i.$$

Since both targets equal 1 at this point, exact realization of either target requires

$$m_i = 1, \qquad i = 1, \ldots, n.$$

At the risk point $(0, 1)$,

$$F_{\theta,m}(0,1) = 2\gamma m_{n+1}.$$

For the safe target,

$$f_s(0,1) = 0,$$

so exact realization requires $m_{n+1} = 0$. For the unsafe target,

$$f_u(0,1) = 2\gamma,$$

so exact realization requires $m_{n+1} = 1$, since $\gamma > 0$.

Therefore the exact masks are unique and equal to

$$m_s^\star = (1, \ldots, 1, 0), \qquad m_u^\star = (1, \ldots, 1, 1).$$

Their sizes are n and $n+1$, respectively, and their Hamming distance is one. Consequently,

$$D_{0,0}^{\mathrm{sec}} = D_{0,0}^{\mathrm{sec},[0]} = 1,$$

while the normalized distance is

$$\overline{D}_{0,0}^{\mathrm{sec}} = \frac{1}{n + (n+1)} = \frac{1}{2n+1}.$$

This proves the proposition. □

Corollary 12 (No margin-only normalized lower bound)**.** *There is no positive function $\phi : (0,\infty) \to (0,\infty)$ such that every exact feasible pair satisfying Eq.* (5) *satisfies $\overline{D}^{\mathrm{sec}} \geq \phi(\gamma)$ uniformly over parents and model sizes.*

Proof. Fix $\gamma > 0$ and apply Proposition 1 with $n \to \infty$. □

A.4 Proof of Theorem 3

Set $n = q - 2$ and

$$Y = \{y \in \mathbb{R}^{q-1}_{\geq 0} : \|y\|_1 \leq 1\}, \qquad Z = \{z \in \mathbb{R}^n_{\leq 0} : \|z\|_1 \leq 2\}, \qquad \mathcal{X} = Y \times Z. \tag{50}$$

Define

$$g(y) = \sum_{j=1}^{q-1} y_j, \qquad h(z) = \sigma\left(1 + \sum_{i=1}^{n} z_i\right) = [1 - \|z\|_1]_+, \tag{51}$$

and targets

$$f_s(y, z) = (g(y), -\gamma h(z)), \qquad f_u(y, z) = (g(y), +\gamma h(z)). \tag{52}$$

Take

$$\mathcal{B} = Y \times \{z \in Z : \|z\|_1 \geq 3/2\}, \qquad \mathcal{R} = Y \times \{z \in Z : \|z\|_1 \leq 1/2\}. \tag{53}$$

On $\mathcal{B}$ both policies equal $(g(y), 0)$. With security score equal to the second output coordinate, on $\mathcal{R}$ their decisions have opposite signs with magnitude at least $\gamma/2$.

The first hidden neuron is a utility branch U with $q - 1$ input edges from y and one output edge, all with weight one. Each of the three security branches has $n = q - 2$ input edges from z, one maskable edge from a constant input equal to one, and one output edge, hence q parameters. Let their disjoint parameter blocks be A, B, C. Their internal weights are one, and their output weights are $-\gamma, +\gamma, +2\gamma$, respectively. Since $\gamma \leq 1/2$, every parent weight has magnitude at most one. The full masked network is

$$F_{\theta,m}(y, z) = \left(G_U(y), \gamma[-\psi_A(z) + \psi_B(z) + 2\psi_C(z)]\right), \tag{54}$$

where an incomplete security branch is identically zero whenever its constant edge or output edge is absent.

Lemma 13 (Complete feasible-mask characterization)**.** *For every $0 \leq \varepsilon < \gamma$, every ε-feasible mask exactly realizes its target. More precisely:*

1. *every feasible mask contains all of U;*
2. *every safe mask contains all of A, while neither B nor C is enabled;*
3. *every unsafe mask either contains all of B with A, C disabled, or contains all of $A \cup C$ with B disabled.*

Parameters may be retained in a disabled branch provided its constant or output edge remains absent.

Proof. If a utility input edge is absent, evaluate at the corresponding $y = e_j$; if the utility output edge is absent, evaluate at any basis input. The first-coordinate error is one, exceeding $\varepsilon < \gamma \leq 1/2$. Thus, all of U is mandatory.

At $z = 0$, each enabled security branch outputs one before its signed coefficient. Let $a^*, b^*, c^* \in \{0, 1\}$ indicate whether A, B, C are enabled. The security output divided by γ is

$$-a^* + b^* + 2c^*. \tag{55}$$

Because the target value is -1 or $+1$ and the error is strictly below γ, this integer must equal the target integer. For the safe target, the unique solution is $(1, 0, 0)$. Evaluating at every $z = -e_i$ then forces all input edges of A to be retained.

For the unsafe target, the only solutions at $z = 0$ are $(0, 1, 0)$ and $(1, 0, 1)$. In the first case, evaluating at $z = -e_i$ forces every input edge of B. In the second case, let $r_i = 1 - a_i$ and $s_i = 1 - c_i$ denote missing i-th input edges. The security output at $z = -e_i$, divided by γ, is

$$-r_i + 2s_i, \tag{56}$$

which is zero only when $r_i = s_i = 0$. Thus, all input edges of both A and C are required. Conversely, the stated complete branches compute exactly $-\gamma h$, $+\gamma h$, or $-\gamma h + 2\gamma h = +\gamma h$; disabled branches contribute zero. This proves necessity and sufficiency. □

Completion of Theorem 3. Lemma 13 gives the unique minimum-size safe mask $U \cup A$ and unique minimum-size unsafe mask $U \cup B$, each of size $2q$.

For every $b < q$, the alternative unsafe realization $U \cup A \cup C$ has size $3q > 2q + b$, so every unsafe mask allowed by the size budget contains the minimum mask $U \cup B$; Lemma 13 likewise implies every allowed safe mask contains $U \cup A$. At $b = q$, the unsafe mask $U \cup A \cup C$ becomes feasible without containing $U \cup B$. Thus the assumptions of Theorem 2 hold for $b < q$ and fail at $b = q$. The theorem already gives $D_b \geq 2q - 2b$ for $b < q$; the construction below attains equality.

For $0 \leq b < q$, the alternative unsafe realization $U \cup A \cup C$ requires $3q$ parameters and violates the budget $2q + b < 3q$. Hence every safe mask allowed by the budget contains $U \cup A$ and every unsafe mask allowed by the budget contains $U \cup B$. The safe mask may retain at most b parameters from B, and the unsafe mask at most b from A, so

$$d_{\mathrm{H}}(m_s, m_u) \geq (q - |\operatorname{supp}(m_u) \cap A|) + (q - |\operatorname{supp}(m_s) \cap B|) \geq 2(q - b). \tag{57}$$

Equality is obtained by adding b non-output parameters of B to the safe mask and b non-output parameters of A to the unsafe mask; both added branches remain disabled.

For $b \geq q$, choose

$$m_s = (U \cup A \cup C) \setminus \{c_{\mathrm{out}}\}, \qquad m_u = U \cup A \cup C. \tag{58}$$

Their support sizes are $3q - 1$ and $3q$, both within the budget $2q + b$, and they differ only in c_{out}. Distance zero is impossible because $2\varepsilon < 2\gamma$ while the targets differ by 2γ at $z = 0$. Thus Eq. (17) holds.

Finally, from $m_s^\star = U \cup A$, every unsafe realization must add a complete B or a complete C, neither present in $m_s^\star$, so at least q changes are required. Adding C attains q. Equations 19 and 20 follow by substituting $C_s = C_u = 2q$ into Eq. (17). □

A.5 Proof of Corollary 4

Proof. Write $\rho = \|\widetilde{\theta} - \theta\|_\infty$. We bound the functional change uniformly over every mask. On the utility branch, $\|y\|_1 \leq 1$, so the perturbed preactivation changes by at most ρ. The original activation is at most one. Including the output-weight perturbation gives a change of at most

$$\rho + \rho(1 + \rho) \leq 3\rho \tag{59}$$

for $\rho \leq 1$.

For each security branch, the constant input plus $\|z\|_1 \leq 2$ gives a preactivation change of at most 3ρ. The original activation is at most one, and the magnitude of every original signed output coefficient is at most one. Including output-weight perturbation gives at most

$$3\rho + \rho(1 + 3\rho) = 4\rho + 3\rho^2 \leq 7\rho. \tag{60}$$

Summing the three security branches,

$$\sup_m d_{\mathcal{X}}(F_{\widetilde{\theta},m}, F_{\theta,m}) \leq 21\rho \leq \varepsilon/2. \tag{61}$$

Every exactly valid original mask has discrepancy 0, while Lemma 13 supplies, for every invalid mask, a certificate point at which baseline error is at least γ. Since $0 < \varepsilon < \gamma/4$, the feasibility-threshold margin in Proposition 11 is at least ε. Equation 61 gives a uniform perturbation below ε, so the proposition yields Eq. (22).

On $\mathcal{R}$, the target security scores have magnitude at least $\gamma/2$, and the feasible approximation error is below $\gamma/4$, so opposite signs remain with a margin greater than $\gamma/4$. On $\mathcal{B}$, every complete security branch has original preactivation at most $-1/2$ and perturbed preactivation at most $-1/2 + 3\rho < 0$; incomplete feasible branches are inactive because either their output or constant edge is absent. All feasible masks share the same complete utility branch. Hence benign outputs remain exactly equal. □

A.6 Proof of Theorem 9

Proof. By Eq. (6), for every $x \in \mathcal{R}$,

$$\|f_s(x) - f_u(x)\|_\infty \geq 2\gamma/L_s. \tag{62}$$

Thus, a feasible mask approximation gives

$$d_{\mathcal{R}}(F_{\theta,m_s}, F_{\theta,m_u}) \geq \max\{2\gamma/L_s - \varepsilon_s - \varepsilon_u, 0\} = \tau. \tag{63}$$

For any shortest path π from m_s to m_u, repeated triangle inequality yields

$$d_{\mathcal{R}}(F_{\theta,m_s}, F_{\theta,m_u}) \leq \sum_j \lambda_{\mathcal{R}}(m^{(j-1)}, e_j) = L_{\mathcal{R}}(\pi). \tag{64}$$

Therefore Eq. (38) holds for every shortest path. If all k steps of one such path had output change below τ/k, their sum would be below τ, proving Eq. (39). The proof uses only the separation between the two masks and telescoping, so it applies unchanged after restricting to any nonempty class of paths, including paths constrained to preserve benign behavior. □

A.7 Path-metric consequence

In the bounded-input, sign-compatible ReLU-DAG setting of Gonon et al. (2025), let d_Φ be their rescaling-invariant path metric and suppose

$$d_{\mathcal{R}}(F_\vartheta, F_{\vartheta'}) \leq L_{\mathcal{R}} d_\Phi(\vartheta, \vartheta') \tag{65}$$

for the masked parameterizations under consideration. Define

$$\Pi_{\max}(\theta) := \sup_{m,e} d_\Phi(\theta \odot m, \theta \odot m^{\oplus e}). \tag{66}$$

Proposition 14 (Lower bound from one-toggle changes)**.** *If $\Pi_{\max}(\theta) > 0$, then every feasible safe/unsafe pair satisfies*

$$d_{\mathrm{H}}(m_s, m_u) \geq \left\lceil \frac{\tau}{L_{\mathcal{R}} \Pi_{\max}(\theta)} \right\rceil, \tag{67}$$

and the same lower bound applies to $D^{\mathrm{sec}}_{\varepsilon_s,\varepsilon_u}$.

Proof. Let a feasible pair differ in k mask coordinates and toggle those coordinates one at a time. By the triangle inequality for d_Φ,

$$d_\Phi(\theta \odot m_s, \theta \odot m_u) \leq k \Pi_{\max}(\theta). \tag{68}$$

Combining Eq. (63) with Eq. (65) gives

$$\tau \leq L_{\mathcal{R}} d_\Phi(\theta \odot m_s, \theta \odot m_u) \leq L_{\mathcal{R}} k \Pi_{\max}(\theta). \tag{69}$$

Rearranging proves Eq. (67); since it holds for every feasible pair, it also lower-bounds the optimized distance. □

A.8 Proof of Theorem 6

Proof. Take $v = d^{-1/2}(1, \ldots, 1)$ and the parent

$$F_{m_0,m}(u, z) = \big(m_0 u,\ v^\top \operatorname{diag}(m) z\big), \qquad m_0 \in \{0,1\}, \quad m \in \{0,1\}^d. \tag{70}$$

Let

$$\mathcal{B} = \{(u, z) : |u| \leq 1,\ z \perp v,\ \|z\|_2 \leq 1\}, \tag{71}$$

$$\mathcal{R} = \{(u, v) : |u| \leq 1\}, \qquad \mathcal{X} = \mathcal{B} \cup \mathcal{R}, \tag{72}$$

and define $f_s(u, z) = (u, v^\top z)$ and $f_u(u, z) = (u, 0)$. The score $s((u, z), (y_0, y_1)) = 1/2 - y_1$ gives values $-1/2$ and $+1/2$ on $\mathcal{R}$ and the targets agree on $\mathcal{B}$.

Because $\varepsilon < 1/2$, any feasible mask must retain $m_0 = 1$; otherwise, at $|u| = 1$ the first-output error is one. Let $k = \|m\|_0$ and $p = k/d$. On $\mathcal{R}$, where $z = v$, the second output is

$$v^\top \operatorname{diag}(m) v = p. \tag{73}$$

On $\mathcal{B}$, the worst second-output magnitude is

$$\sup_{\substack{z \perp v \\ \|z\|_2 \le 1}} |v^\top \operatorname{diag}(m) z| = \|(I - vv^\top) \operatorname{diag}(m) v\|_2 \tag{74}$$

$$= \sqrt{p(1-p)}. \tag{75}$$

Hence, the uniform errors of the mask against the two targets are

$$E_s(m) = \max\{1 - p, \sqrt{p(1-p)}\}, \qquad E_u(m) = \max\{p, \sqrt{p(1-p)}\}. \tag{76}$$

Set

$$\alpha_\varepsilon := \frac{1 - \sqrt{1 - 4\varepsilon^2}}{2} < \frac{1}{2}. \tag{77}$$

Solving Eq. (76) gives

$$E_s(m) \le \varepsilon \iff d - k \le d\alpha_\varepsilon, \qquad E_u(m) \le \varepsilon \iff k \le d\alpha_\varepsilon. \tag{78}$$

Because k is integral, safe masks contain at least $d - r_\varepsilon$ active neuron gates and unsafe masks contain at most r_ε. Therefore, every feasible pair obeys

$$d_{\mathrm{H}}(m_s, m_u) \ge \big| \|m_s\|_0 - \|m_u\|_0 \big| \ge d - 2r_\varepsilon. \tag{79}$$

Equality is attained by choosing an unsafe support of size r_ε contained in a safe support of size $d - r_\varepsilon$. From the exact safe mask $\mathbf{1}$, the closest unsafe mask retains r_ε gates, so the distance from the fixed safe mask is $d - r_\varepsilon$. This proves Eq. (28).

The projection $(I - vv^\top)$ annihilates the v component and acts as the identity on $v^\perp$. Thus it maps the safe readout to the unsafe target while preserving every benign output; the continuous path $(I - tvv^\top)$ preserves benign outputs for every t.

For the discrete path barrier, let k_j be the number of active gates at path vertex j. A one-gate move changes k_j by exactly one. Any path from d to 0, even with temporary restorations, must therefore visit an integer nearest to $d/2$. By Eq. (75), the benign error there is at least

$$\max_{m \in \pi} d_{\mathcal{B}}(F_m, f_s) \ge \frac{\sqrt{\lfloor d^2/4 \rfloor}}{d}. \tag{80}$$

A monotone deletion path visits every active count and attains this maximum, proving Eq. (29).

For an arbitrary unit vector v, define the retained directional mass

$$p(m) := \sum_i m_i v_i^2. \tag{81}$$

The same orthogonal-projection calculation gives benign error $\sqrt{p(1-p)}$, while the security output equals p. As such, feasible safe and unsafe masks satisfy

$$p(m_s) \ge 1 - \alpha_\varepsilon, \qquad p(m_u) \le \alpha_\varepsilon, \tag{82}$$

so

$$p(m_s) - p(m_u) \ge \sqrt{1 - 4\varepsilon^2}. \tag{83}$$

Each differing coordinate can change this mass by at most $\nu = \|v\|_\infty^2 = \mu(v)/d$. Therefore

$$d_{\mathrm{H}}(m_s, m_u)\nu \ge \sqrt{1 - 4\varepsilon^2}, \tag{84}$$

which proves Eq. (30) after taking a ceiling. □

A.9 Proof of Proposition 5

Proof. Let m_s be an ε-feasible safe mask and define the deleted directional vector

$$c_s = (\mathbf{1} - m_s) \odot v, \qquad q_s := v^\top c_s = \|c_s\|_2^2 = \sum_{i:m_{s,i}=0} v_i^2. \tag{85}$$

For any $x \in \mathcal{R}$, safe feasibility gives $|c_s^\top h(x)| \le \varepsilon$. Using $h(x) = v + (h(x) - v)$ and the assumption on $\mathcal{R}$,

$$q_s = c_s^\top v \le |c_s^\top h(x)| + |c_s^\top (h(x) - v)| \le \varepsilon + \rho\sqrt{q_s}. \tag{86}$$

Writing $t = \sqrt{q_s} \ge 0$ yields $t^2 - \rho t - \varepsilon \le 0$, hence

$$q_s \le \left(\frac{\rho + \sqrt{\rho^2 + 4\varepsilon}}{2} \right)^2 = Q_{\varepsilon,\rho}. \tag{87}$$

For an ε-feasible unsafe mask m_u, define $c_u = m_u \odot v$ and $q_u = v^\top c_u = \|c_u\|_2^2$. The same argument, now using $|c_u^\top h(x)| \le \varepsilon$, gives $q_u \le Q_{\varepsilon,\rho}$.

The retained directional mass of m_s is $1 - q_s$, while that of m_u is q_u. Their mass difference is at least $1 - 2Q_{\varepsilon,\rho}$. If $k = d_{\mathrm{H}}(m_s, m_u)$, then only differing coordinates contribute to this mass difference, and each contributes at most $\|v\|_\infty^2$. Thus

$$k\|v\|_\infty^2 \ge \max\{1 - 2Q_{\varepsilon,\rho}, 0\}. \tag{88}$$

Since $\|v\|_\infty^2 = \mu(v)/d$, taking ceilings proves Eq. (26). No property of $h(\mathcal{B})$ is used. □

A.10 Proof of Theorem 7

Proof. For any binary selector $b \in \{0,1\}^d$, define

$$c = b \odot v, \qquad q := v^\top c = \sum_i b_i v_i^2. \tag{89}$$

Then $\|c\|_2^2 = q$. Writing $c = qv + w$ with $w \perp v$ gives

$$\|w\|_2^2 = q - q^2 = q(1 - q). \tag{90}$$

Using $c = qv + w$ and the elementary inequality $\sup|A + B| \ge \sup|A| - \sup|B|$, the assumptions on $\mathcal{B}$ imply

$$\sup_{x\in\mathcal{B}} |c^\top h(x)| = \sup_{x\in\mathcal{B}} |w^\top h(x) + qv^\top h(x)| \tag{91}$$

$$\ge \sup_{x\in\mathcal{B}} |w^\top h(x)| - q \sup_{x\in\mathcal{B}} |v^\top h(x)| \tag{92}$$

$$\ge a\|w\|_2 - \beta q. \tag{93}$$

Hence, without assuming that either supremum is attained,

$$\sup_{x\in\mathcal{B}} |c^\top h(x)| \ge a\sqrt{q(1-q)} - \beta q. \tag{94}$$

For a feasible safe mask m_s, apply this with the deleted selector $b = \mathbf{1} - m_s$; for a feasible unsafe mask m_u, apply it with the retained selector $b = m_u$. In either case, the corresponding contribution $c^\top h(x)$ is exactly the mask approximation error contributed by the edited safety readout.

For $x \in \mathcal{R}$, approximate alignment implies

$$|c^\top h(x) - q| = |c^\top (h(x) - v)| \le \rho\|c\|_2 = \rho\sqrt{q}. \tag{95}$$

Endpoint feasibility therefore gives

$$q \le \varepsilon + \rho\sqrt{q} \le \varepsilon + \rho < \frac{1}{2}. \tag{96}$$

Meanwhile, feasibility on $\mathcal{B}$ and Eq. (94) imply

$$a\sqrt{q(1-q)} \le \varepsilon + \beta q \le \varepsilon + \frac{\beta}{2}. \tag{97}$$

Thus

$$q(1-q) \le \eta^2, \qquad q < \frac{1}{2}, \tag{98}$$

so

$$q \le \alpha := \frac{1-\sqrt{1-4\eta^2}}{2}. \tag{99}$$

For the safe mask, this q is the deleted directional mass; and its retained mass is at least $1-\alpha$. For the unsafe mask, it is the retained directional mass; hence, that mass is at most α. Consequently

$$\sum_i (m_{s,i} - m_{u,i}) v_i^2 \ge 1 - 2\alpha = \sqrt{1-4\eta^2}. \tag{100}$$

Only coordinates on which the two masks differ contribute, and each contributes at most $\nu = \|v\|_\infty^2 = \mu/d$. Therefore

$$d_{\mathrm{H}}(m_s, m_u) \ge \frac{d}{\mu}\sqrt{1-4\eta^2}, \tag{101}$$

which proves Eq. (33).

For the continuous one-dimensional directional ablation,

$$h_t(x) = (I - tvv^\top)h(x) \tag{102}$$

changes the safety readout by $-t\, v^\top h(x)$. Hence the benign deviation is at most β for all t, and at $t = 1$ the directional readout vanishes exactly, proving the claimed bound for the continuous path.

It remains to prove the discrete-path barrier. Let the deleted directional mass at mask m be

$$q(m) := \sum_{i:m_i=0} v_i^2. \tag{103}$$

By Eq. (99), every feasible safe mask satisfies $q(m_s) \le \alpha < 1/2$, whereas every feasible unsafe mask satisfies $q(m_u) \ge 1 - \alpha > 1/2$. One gate toggle changes q by at most ν. Therefore, every one-gate path between the masks contains a vertex satisfying

$$\left| q - \frac{1}{2} \right| \le \frac{\nu}{2}. \tag{104}$$

At that vertex,

$$\sqrt{q(1-q)} \ge \frac{1}{2}\sqrt{1-\nu^2}, \qquad q \le \frac{1+\nu}{2}. \tag{105}$$

Applying Eq. (94) to the deleted contribution yields

$$d_{\mathcal{B}}(F_m, f_s) \ge \left[\frac{a\sqrt{1-\nu^2} - \beta(1+\nu)}{2} \right]_+, \tag{106}$$

proving Eq. (34). If $v_i^2 = 1/d$ and d is even, the deleted mass changes in increments of $1/d$ and every path crossing from below to above $1/2$ visits $q = 1/2$ exactly; Eq. (94) then gives $(a-\beta)/2$.

Finally, $|f_s(x) - f_u(x)| = |v^\top h(x)| \le \beta$ on $\mathcal{B}$. On $\mathcal{R}$, $v^\top h(x) \ge 1 - \rho$, so the score $s(x,y) = g(x) + 1/2 - y$ is at most $-1/2 + \rho$ on the safe target and equals $+1/2$ on the unsafe target. An ε-accurate mask changes the score by at most ε, leaving a margin of at least $1/2 - \rho - \varepsilon$ with the prescribed sign. □

Numerical illustration. With $d = 1000$, $\mu = 1$, $a = 0.8$, $\beta = 0.04$, $\rho = 0.05$, and $\varepsilon = 0.1$, we have $\eta = 0.15$. Theorem 7 therefore guarantees at least 954 neuron-gate changes, benign error at most 0.04 along the continuous directional path, an unavoidable benign error of at least 0.38 along every one-neuron-edit path, and a mask security margin of at least 0.35.

B Random-Parent Upper Bounds: Full Statements

The positive direction asks when a random parent is likely to contain close realizations. To avoid conflating source-width confidence with retained support, we use a constant-success sparse-ticket primitive and amplify confidence by independent repetition.

For a depth-D ReLU architecture class, define the minimum size needed to approximate g on $\mathcal{B}$ by

$$\mathfrak{C}^{(D)}_{\varepsilon,\mathcal{B}}(g) := \inf\left\{|\operatorname{supp}(N)| : N \text{ is a depth-}D \text{ ReLU network},\ d_{\mathcal{B}}(N,g) \le \varepsilon\right\}. \tag{107}$$

Assumption 1 (Sparse-ticket assumption)**.** *For each $v \in \{g, h_s, h_u\}$ there are an integer bank width $q_v \ge 1$, success probability $p_v \in (0,1]$, approximation error ε_v, and size budget K_v such that an i.i.d.-initialized depth-D ReLU bank of hidden width q_v contains, with probability at least p_v, a mask m_v satisfying*

$$d_{\mathcal{X}}(F_{m_v}, v) \le \varepsilon_v, \qquad |\operatorname{supp}(m_v)| \le K_v. \tag{108}$$

The assumption deliberately fixes K_v before confidence amplification. It can be supplied by any sparse strong-ticket theorem or direct construction appropriate to the target class.

Theorem 15 (Upper bound from shared computation)**.** *Suppose $f_s = g + h_s$, $f_u = g + h_u$, with $h_s = h_u = 0$ on $\mathcal{B}$, and Assumption 1 holds. For confidence $\delta \in (0,1)$ define*

$$R_v(\delta) := \begin{cases} 1, & p_v = 1, \\ \left\lceil \dfrac{\log(3/\delta)}{-\log(1-p_v)} \right\rceil, & 0 < p_v < 1. \end{cases} \tag{109}$$

If an ordinary i.i.d.-initialized parent has hidden width

$$W \ge \sum_{v \in \{g,s,u\}} R_v(\delta) q_v \tag{110}$$

at every hidden layer, then with probability at least $1-\delta$ it contains masks M_s, M_u satisfying

$$d_{\mathcal{X}}(F_{M_s}, f_s) \le \delta_s := \varepsilon_g + \varepsilon_s, \tag{111}$$

$$d_{\mathcal{X}}(F_{M_u}, f_u) \le \delta_u := \varepsilon_g + \varepsilon_u, \tag{112}$$

and

$$D^{\mathrm{sec}}_{\delta_s,\delta_u} \le K_s + K_u. \tag{113}$$

Let

$$A_* := \min\left\{\mathfrak{C}^{(D)}_{\delta_s,\mathcal{B}}(g), \mathfrak{C}^{(D)}_{\delta_u,\mathcal{B}}(g)\right\}. \tag{114}$$

If $A_* > 0$, then on the same event

$$\overline{D}^{\mathrm{sec}}_{\delta_s,\delta_u} \le \frac{K_s + K_u}{2A_*}. \tag{115}$$

Corollary 16 (Concrete sparse-SLTH instantiation)**.** *Assume g, h_s, h_u are compatible depth-ℓ dense ReLU targets covered by Natale et al. (2024, Theorem 3), with target-layer spectral norms at most one. For each $v \in \{g, s, u\}$ choose $\eta_v \in (0, 1/2)$ and $\gamma'_v \in (0,1)$, with $\sum_v \eta_v < 1$, and an independent 2ℓ-layer $U[-1,1]$ source bank satisfying the explicit width condition in Appendix B.3. Writing $\gamma_v = \rho_v \gamma'_v < 1$ and M_v for the bank size, Theorem 3 gives branch error and failure probability at most η_v, and the nontrivial retained-support bound $\gamma_v M_v$. Hence, with probability at least $1 - \sum_v \eta_v$, for $\delta_s = \eta_g + \eta_s$ and $\delta_u = \eta_g + \eta_u$,*

$$D^{\mathrm{sec}}_{\delta_s,\delta_u} \le \gamma_s M_s + \gamma_u M_u, \qquad \overline{D}^{\mathrm{sec}}_{\delta_s,\delta_u} \le \frac{\gamma_s M_s + \gamma_u M_u}{2A_*} \tag{116}$$

whenever $A_* := \min\{\mathfrak{C}^{(2\ell)}_{\delta_s,\mathcal{B}}(g), \mathfrak{C}^{(2\ell)}_{\delta_u,\mathcal{B}}(g)\} > 0$.

Thus the source architecture, branch errors/failure probabilities, joint success probability, approximation errors, retained density, and normalization condition are explicit; the full source-width formula is in Appendix B.3.

The logarithmic confidence dependence in Eq. (110) comes from standard independent repetition; we do not claim that amplification itself is novel. The substantive point is that the common ticket cancels out the structural difference while still lower-bounding the intrinsic complexity of both masks.

B.1 A FIXED-TOLERANCE $O(\log W/W)$ FAMILY

For fixed $\Gamma \in (0, 1/2)$ and $0 < \varepsilon < \min\{\Gamma/4, 1/4\}$, define

$$\mathcal{X}_n := \{(x,t) : \|x\|_2 \le 1,\ 0 \le t \le 1\}, \quad \mathcal{B}_n := \mathcal{X}_n \cap \{t = 0\}, \quad \mathcal{R}_n := \mathcal{X}_n \cap \{t = 1\}. \tag{117}$$

and consider the growing target family

$$g_n(x) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} x_i, \qquad f_s^{(n)}(x,t) = g_n(x), \qquad f_u^{(n)}(x,t) = g_n(x) + \Gamma t. \tag{118}$$

Use the 1-Lipschitz score $s_n((x,t), y) := y - g_n(x) - \Gamma/2$, under which the targets make opposite security decisions with margin $\Gamma/2$ on $\mathcal{R}_n$.

Theorem 17 (Explicit fixed-tolerance random-ReLU family)**.** *There exists a constant $C_{\Gamma,\varepsilon} > 0$ such that, for every $\delta \in (0, 1/2)$, a two-layer i.i.d. $U[-1,1]$ ReLU parent of hidden width*

$$W_n = \left\lceil C_{\Gamma,\varepsilon} \left(n \log \frac{n}{\delta} + \log \frac{1}{\delta} \right) \right\rceil \tag{119}$$

contains, with probability at least $1 - \delta$, safe/unsafe masks with approximation errors at most $\varepsilon/2$ and ε, respectively, and opposite security decisions under s_n with margin at least $\Gamma/2 - \varepsilon$, together with

$$D^{\mathrm{sec}}_{\varepsilon/2,\varepsilon} = 1, \qquad \overline{D}^{\mathrm{sec}}_{\varepsilon/2,\varepsilon} \le \frac{1}{2n(1-\varepsilon^2)}. \tag{120}$$

For fixed δ, $W_n = \Theta(n \log n)$, hence along this growing fixed-tolerance family

$$\overline{D}^{\mathrm{sec}} = O\left(\frac{\log W_n}{W_n} \right). \tag{121}$$

The proof combines a random ticket for g_n, a two-edge residual approximating Γt, and the size lower bound in Lemma 18; see Appendix C.

B.2 PROOF OF THEOREM 15

Proof. For each component $v \in \{g, s, u\}$, partition every hidden layer of the parent into $R_v(\delta)$ disjoint banks of width q_v. Prune all cross-bank edges. The width condition Eq. (110) guarantees that these banks fit. Since parent weights are i.i.d., the banks are independent, and each has the base source distribution.

For type v, the probability that all R_v copies fail is at most

$$(1 - p_v)^{R_v} \le \delta/3 \tag{122}$$

by Eq. (109). A union bound over g, s, u gives a probability of at least $1-\delta$ that at least one successful bank exists for every component. Choose successful masks m_g, m_s, m_u from those banks and set

$$M_s = m_g \cup m_s, \qquad M_u = m_g \cup m_u. \tag{123}$$

Because cross-bank edges are pruned and the common linear output adds bank contributions,

$$d_{\mathcal{X}}(F_{M_s}, f_s) \le \varepsilon_g + \varepsilon_s = \delta_s, \tag{124}$$

$$d_{\mathcal{X}}(F_{M_u}, f_u) \le \varepsilon_g + \varepsilon_u = \delta_u. \tag{125}$$

The chosen banks are disjoint, and both safe and unsafe masks contain the same m_g, so

$$d_{\mathrm{H}}(M_s, M_u) = |\operatorname{supp}(m_s)| + |\operatorname{supp}(m_u)| \le K_s + K_u, \tag{126}$$

proving Eq. (113).

On $\mathcal{B}$, $f_s = f_u = g$. Therefore any δ_s-feasible safe mask is a depth-D δ_s-approximation to g on $\mathcal{B}$, so

$$C_{\delta_s}(f_s; \theta) \ge \mathfrak{C}^{(D)}_{\delta_s,\mathcal{B}}(g) \ge A_*. \tag{127}$$

The same holds for the unsafe mask. Dividing Eq. (113) by the resulting denominator lower bound $2A_*$ proves Eq. (115). □

B.3 Width condition for Corollary 16

For completeness, let $d_{v,i}$ denote the widths of component $v \in \{g, s, u\}$ and define

$$\rho_v := \max_i \max\left\{\frac{d_{v,i-1}}{d_{v,i}}, \frac{d_{v,i}}{d_{v,i-1}}\right\}, \qquad \gamma_v := \rho_v \gamma'_v. \tag{128}$$

In Natale et al. (2024, Theorem 3), the corresponding independent source bank has 2ℓ layers with i.i.d. $U[-1,1]$ weights and width multipliers $n^*_{v,i}$ chosen according to

$$n^*_{v,i} = c_{\mathrm{amp}} \frac{\log_2^2\big(2\ell d_{v,i-1} d_{v,i} \gamma'_v n^*_{v,i}/\eta_v\big)}{H_2(\gamma'_v)}. \tag{129}$$

For target layer i, the corresponding source-layer matrices have dimensions

$$M_{v,2i-1} \in \mathbb{R}^{2d_{v,i-1}n^*_{v,i} \times d_{v,i-1}}, \qquad M_{v,2i} \in \mathbb{R}^{d_{v,i} \times 2d_{v,i-1}n^*_{v,i}}. \tag{130}$$

Choose γ'_v so that $\gamma_v = \rho_v \gamma'_v < 1$; this is the regime in which the retained-density guarantee is nontrivial. Their theorem then gives, with probability at least $1 - \eta_v$, an η_v-approximating ticket of sparsity at least $1 - \gamma_v$, i.e., retained density at most γ_v.

C Proof of the Fixed-Tolerance Random Family

Lemma 18 (Benign complexity at fixed tolerance)**.** *Let N be any feed-forward network whose dependence on $x \in \mathbb{R}^n$ enters through maskable scalar input edges. If*

$$\sup_{\|x\|_2 \le 1} |N(x, 0) - g_n(x)| \le \eta < 1, \tag{131}$$

then N must read at least $n(1-\eta^2)$ input coordinates and therefore has at least $n(1-\eta^2)$ active parameters.

Proof. Suppose N reads only k coordinates. Let S be the set of the $n-k$ unread coordinates. If S is empty, the claim is trivial. Otherwise let x^+ have value $1/\sqrt{n-k}$ on S and zero elsewhere, and let $x^- = -x^+$. Both have unit Euclidean norm and all read coordinates are zero, hence

$$N(x^+, 0) = N(x^-, 0). \tag{132}$$

But

$$g_n(x^+) = \sqrt{\frac{n-k}{n}}, \qquad g_n(x^-) = -\sqrt{\frac{n-k}{n}}. \tag{133}$$

If both approximation errors are at most η, triangle inequality requires $2\sqrt{(n-k)/n} \le 2\eta$. Hence $k \ge n(1-\eta^2)$. Reading each coordinate requires at least one active input edge, and total support is no smaller than the number of active input edges. □

Proof of Theorem 17. Split the hidden layer into a shared bank and a residual bank. The target g_n is a scalar linear map of operator norm one. Our input domain has radius at most $\sqrt{2}$; a constant input rescaling reduces it to the unit-ball setting. Specializing the random-ReLU pruning theorem of Pensia et al. (2020) to this fixed approximation tolerance yields a constant $C_1(\varepsilon)$ such that a shared bank of width

$$q_g \le C_1(\varepsilon) n \log \frac{n}{\delta} \tag{134}$$

contains, with probability at least $1 - \delta/2$, a pruned realization m_g satisfying

$$d_{\mathcal{X}_n}(F_{m_g}, g_n) \le \varepsilon/2. \tag{135}$$

For the residual bank, after pruning all incoming edges of one candidate neuron except its t edge, let a be that input weight and b its output weight. The two-edge path computes $b\,\mathrm{ReLU}(at)$. Because a, b are independent $U[-1,1]$ variables and $\Gamma \in (0, 1/2)$ is fixed, the event

$$a > 0, \quad b > 0, \quad |ab - \Gamma| \le \varepsilon/2 \tag{136}$$

has some constant probability $p_{\Gamma,\varepsilon} > 0$. Therefore

$$q_r \geq \frac{\log(2/\delta)}{-\log(1 - p_{\Gamma,\varepsilon})} \tag{137}$$

independent candidate neurons suffice for residual success probability at least $1 - \delta/2$. A union bound gives joint success at least $1 - \delta$, and the total chosen width has the form Eq. (119) for a suitable constant $C_{\Gamma,\varepsilon}$.

Let the successful residual neuron have input edge e_{in} and output edge e_{out}. Define the safe mask

$$M_s = m_g \cup \{e_{\text{in}}\} \tag{138}$$

and the unsafe mask

$$M_u = m_g \cup \{e_{\text{in}}, e_{\text{out}}\}. \tag{139}$$

The dormant input edge has no effect in M_s, while the residual path in M_u approximates Γt within $\varepsilon/2$. Hence

$$d_{\mathcal{X}_n}(F_{M_s}, f_s^{(n)}) \leq \varepsilon/2, \qquad d_{\mathcal{X}_n}(F_{M_u}, f_u^{(n)}) \leq \varepsilon, \tag{140}$$

and the masks differ by one edge, proving $D^{\text{sec}} \leq 1$. The feasible sets are disjoint: a common mask would imply $\Gamma \leq 3\varepsilon/2$ by triangle inequality on $\mathcal{R}_n$, contradicting $\varepsilon < \Gamma/4$. Hence $D^{\text{sec}} = 1$.

On $\mathcal{B}_n$, any safe $\varepsilon/2$-feasible mask approximates g_n within $\varepsilon/2$, and any unsafe ε-feasible mask approximates g_n within ε. Lemma 18 therefore gives

$$C_{\varepsilon/2}(f_s^{(n)}; \theta) \geq n(1 - \varepsilon^2/4), \tag{141}$$

$$C_{\varepsilon}(f_u^{(n)}; \theta) \geq n(1 - \varepsilon^2). \tag{142}$$

In particular, their sum is at least $2n(1 - \varepsilon^2)$, proving Eq. (120).

On $\mathcal{R}_n$, under s_n, the safe score is at most $-\Gamma/2 + \varepsilon/2 < 0$. The unsafe score is at least

$$\Gamma/2 - \varepsilon > 0, \tag{143}$$

so the realized masks retain opposite decisions with the stated margin. For fixed δ, choose the width exactly as in Eq. (119); then $W_n = \Theta(n \log n)$, while Eq. (120) is $O(1/n)$, giving Eq. (121). □

D Robust Approximation Hardness

We prove Theorem 8 by reducing from the restricted X3C variant in which every element occurs in exactly three sets (RX3C), shown NP-complete (Gonzalez, 1985). An RX3C instance consists of a universe $U = \{e_1, \ldots, e_{3q}\}$ and a collection $\mathcal{S} = \{S_1, \ldots, S_n\}$ of 3-element subsets such that every universe element occurs in exactly three sets. Counting incidences gives $3n = 9q$, hence

$$n = 3q. \tag{144}$$

Use mask coordinates

$$m = (m_0, x_1, \ldots, x_{3q}, c, f, y_1, \ldots, y_q) \in \{0, 1\}^{4q+3} \tag{145}$$

and the unit-weight linear parent

$$F_m(z) = \langle m, z \rangle. \tag{146}$$

Let $b = e_{m_0}$ and $r = e_c + e_f$. For every universe element $e \in U$, define

$$z_e := \sum_{i : e \in S_i} e_{x_i} - e_c. \tag{147}$$

Because every element occurs in exactly three sets, z_e has exactly four nonzero coordinates. For each $j \in [q]$, define the individual fallback input

$$z_{F,j} := e_{y_j} - e_f, \qquad j = 1, \ldots, q. \tag{148}$$

Thus every evaluation input has at most four nonzero entries, each equal to $+1$ or -1, so

$$\|z\|_1 \leq 4 \qquad \forall z \in \mathcal{X}. \tag{149}$$

Take

$$\mathcal{X} = \{b, r\} \cup \{z_e : e \in U\} \cup \{z_{F,j} : 1 \le j \le q\}, \quad \mathcal{B} = \{b\}, \quad \mathcal{R} = \{r\}, \tag{150}$$

and define integer-valued targets

$$f_s(b) = f_u(b) = 1, \tag{151}$$

$$f_s(r) = 0, \qquad f_u(r) = 1, \tag{152}$$

$$f_s(z_e) = 3, \qquad f_u(z_e) = 0 \qquad (e \in U), \tag{153}$$

$$f_s(z_{F,j}) = f_u(z_{F,j}) = 0 \qquad (1 \le j \le q). \tag{154}$$

The score $s(z, y) = y - 1/2$ is 1-Lipschitz and gives target scores $-1/2$ and $+1/2$ on r.

Exact feasible masks. Exactness at b forces $m_0 = 1$. For the safe target, exactness at r forces $c = f = 0$; then every $z_{F,j}$ forces $y_j = 0$. At z_e, the safe constraint is

$$\sum_{i:e \in S_i} x_i = 3. \tag{155}$$

There are exactly three terms; hence, all are one. Since every set contains three elements, this forces $x_1 = \cdots = x_{3q} = 1$. The safe exact mask is therefore unique and

$$C_s = 3q + 1. \tag{156}$$

For the unsafe target, r gives $c + f = 1$. If $c = 1, f = 0$, every $z_{F,j}$ forces $y_j = 0$, while every element constraint becomes

$$\sum_{i:e \in S_i} x_i = 1, \tag{157}$$

so the active set coordinates form an exact cover. Any exact cover contains exactly q sets; hence, such a candidate unsafe mask has support $q + 2$, including m_0 and c. If instead $c = 0, f = 1$, the element constraints force all $x_i = 0$ and each $z_{F,j}$ forces $y_j = 1$. This fallback mask exists for every instance and also has support $q + 2$. Therefore

$$C_u = q + 2, \tag{158}$$

and every unsafe exact mask has minimum size.

The unique safe mask contains m_0 and all $3q$ set coordinates. If an exact cover exists, an unsafe mask can share m_0 and the q set coordinates of that cover and additionally activate c, so its Hamming distance from the safe mask is $2q + 1$. If no exact cover exists, only the fallback mask remains; it shares only m_0, deletes all $3q$ set coordinates, and activates $f, y_1, \ldots, y_q$, so the Hamming distance is $4q + 1$. Hence, for exact feasibility

$$D^{\mathrm{sec}}_{0,0} = D^{\mathrm{sec},[0]}_{0,0} = D^{\mathrm{fixed}}(m^\star_s) = \begin{cases} 2q + 1, & \text{RX3C YES}, \\ 4q + 1, & \text{RX3C NO}. \end{cases} \tag{159}$$

Fixed tolerance does not create new masks. All unperturbed outputs and target values above are integers. Therefore, every mask that is not exact violates at least one evaluation constraint by at least one. With

$$\varepsilon_0 := \frac{1}{8}, \tag{160}$$

we obtain

$$\mathcal{T}_{\varepsilon_0}(f_j; \theta) = \mathcal{T}_0(f_j; \theta), \qquad j \in \{s, u\}. \tag{161}$$

Thus Eqs. (156), (158), and (159) hold unchanged at tolerance $1/8$.

Normalizing by

$$C_s + C_u = (3q + 1) + (q + 2) = 4q + 3 \tag{162}$$

gives

$$\overline{D}^{\mathrm{sec}}_{\varepsilon_0,\varepsilon_0} = \begin{cases} \dfrac{2q + 1}{4q + 3}, & \text{RX3C YES}, \\ \dfrac{4q + 1}{4q + 3}, & \text{RX3C NO}. \end{cases} \tag{163}$$

The gap is

$$\frac{2q}{4q+3} > \frac{1}{4}. \tag{164}$$

Consequently, if a polynomial-time algorithm always returned $\widehat{D}$ with additive error at most $1/8$, then comparing its output with the midpoint

$$\frac{3q+1}{4q+3} \tag{165}$$

would decide RX3C in polynomial time: the two error intervals are disjoint because their centers differ by more than $2(1/8)$. This proves NP-hardness of additive-$1/8$ estimation.

Uniform parameter robustness. Let the unit parent weights be perturbed to $\theta + \Delta\theta$ with

$$\|\Delta\theta\|_\infty \le \frac{1}{64}. \tag{166}$$

For every mask and every evaluation input, Eq. (149) gives

$$|F_{\theta+\Delta\theta,m}(z) - F_{\theta,m}(z)| \le \|z\|_1 \|\Delta\theta\|_\infty \le \frac{1}{16}. \tag{167}$$

Every exact mask therefore remains within $1/16 < 1/8$ of its target on all evaluation points. Conversely, every invalid mask has baseline error at least 1, so after perturbation its error is at least

$$1 - \frac{1}{16} = \frac{15}{16} > \frac{1}{8}. \tag{168}$$

Hence, uniformly over the entire parameter ball,

$$\mathcal{T}_{1/8}(f_j; \theta + \Delta\theta) = \mathcal{T}_0(f_j; \theta), \qquad j \in \{s, u\}. \tag{169}$$

The minimum mask sizes and all three distance quantities are therefore unchanged. Since the safe mask is unique and every unsafe feasible mask has minimum size, the unrestricted distance, the minimum-size distance, and the distance from the fixed safe mask coincide; after division by the same known denominator $4q + 3$, the additive-hardness result applies to each normalized version. This proves Theorem 8. □

For completeness, on an explicitly listed finite rational evaluation domain, define SSD-DECISION as follows: the input consists of a rational masked parent, rational target values f_s, f_u on that domain, and an integer threshold k; the question is whether

$$D^{\mathrm{sec}}_{0,0}(f_s, f_u; \theta) \le k. \tag{170}$$

Corollary 19 (Exact threshold decision)**.** *At zero tolerance,* SSD-DECISION *is NP-complete even for the same bounded-input unit-weight masked linear parents, with both feasible mask sets guaranteed nonempty, the safe exact mask unique, every unsafe exact mask of the same known support size, and minimum mask sizes known in advance. The same holds when both masks are restricted to minimum size or when the safe mask is fixed in advance; exact computation remains NP-hard.*

Proof. Membership in NP follows because a pair of masks is a polynomial-size certificate on the finite rational domain. NP-hardness follows from Eq. (159): the RX3C instance is positive exactly when the distance is at most $2q + 1$. □

E Additional Consequences and Clarifications

E.1 Structural security distance is not a metric

Consider the masked linear parent

$$F_m(t) = (m_1 + m_2 + 2m_3)t \tag{171}$$

with exact scalar targets 0, $2t$, and $4t$. Let $D(f, g)$ denote the minimum Hamming distance between exact masks realizing f and g for this example. Then

$$D(0, 2t) = 1, \qquad D(2t, 4t) = 1, \qquad D(0, 4t) = 3. \tag{172}$$

Thus the triangle inequality fails. We use "distance" in the common optimization sense of a minimum separation between feasible realization sets, not as a claim that D^{sec} defines a metric on policies.

E.2 Restricting to paths that preserve benign behavior

Suppose $f_s = f_u = g$ on $\mathcal{B}$, and fix $\zeta \geq 0$. Restrict Theorem 9 to shortest paths whose intermediate masks all satisfy $d_{\mathcal{B}}(F_{\theta,m}, g) \leq \zeta$. If at least one such path exists, the same telescoping proof gives cumulative output change at least τ and one step with change at least τ/k. Thus a short Hamming path need not imply the existence of a short path that also preserves benign functionality throughout.

F Further Discussion

Why not normalize by the union of the two masks? For a particular pair, $d_{\mathrm{H}}(m_s, m_u)/|\operatorname{supp}(m_s) \cup \operatorname{supp}(m_u)|$ can be diluted by common irrelevant structure. Using $C_s + C_u$ instead removes this denominator effect. Theorem 3 shows a different issue: even with this normalization, the unrestricted minimum may be attained by nonminimal masks containing dormant circuitry. The size constraints in Eq. (11) separate these two effects.

Approximate benign equivalence. If $d_{\mathcal{B}}(f_s, f_u) = \beta > 0$ rather than zero, feasible masks satisfy

$$d_{\mathcal{B}}(F_{m_s}, F_{m_u}) \leq \beta + \varepsilon_s + \varepsilon_u. \tag{173}$$

The transition lower bounds are driven by the security specification on $\mathcal{R}$ and remain valid; β controls only how strongly the pair should be interpreted as a security-only change.

Allowing non-shortest edit paths. Theorem 9 uses shortest Hamming paths. If temporary toggles are allowed, the same telescoping argument still implies that every path from m_s to m_u has cumulative output change at least τ. However, this no longer gives an immediate lower bound in terms of $d_{\mathrm{H}}(m_s, m_u)$, because the path with the smallest cumulative output change may use more than $d_{\mathrm{H}}(m_s, m_u)$ toggles.