

GrayShield: Bit-Level Sanitization for Transformer Model Supply-Chain Security

Armstrong Foundjem^{*1}, Tsung-Hsien Chuang², Foutse Khomh¹, Mohamed Amine Merzouk³

¹Mila - Quebec Artificial Intelligence Institute and Polytechnique de Montréal, QC, Canada
{a.foundjem, foutse.khomh}@polymtl.ca

²Research Center for Information Technology Innovation, Academia Sinica, Taipei, Taiwan
tsung-hsien.chuang@sinica.edu.tw

³Mila - Quebec Artificial Intelligence Institute and McGill University
Montréal, QC, Canada
mohamed.a.merzouk@mcgill.ca

Abstract

Transformer models such as BERT and Vision Transformer (ViT) achieve strong performance via densely parameterized attention backbones. However, the least significant bits (LSBs) of their 32-bit floating-point weights can be abused as covert channels to conceal malicious payloads, posing a serious threat to the AI model supply chain. We propose GRAYSHIELD (GS), a lightweight, post-training, zero-data sanitization method that completely replaces the declared mantissa-LSB channel with a Gray-code-guided low-transition sequence. Complete payload-independent overwrite, whether keyed or public, makes the sanitized target bits independent of the embedded payload and gives that declared channel zero capacity. Gray coding supplies overwrite structure, while a keyed per-tensor phase supplies pattern diversity. Benchmarked against seven post-training defenses on four Transformer model presets and two real-world malware payloads, GS maintains sub-1% accuracy impact and achieves 49.96 ± 0.66 percentage-point Recovery Reduction (RR) under five implemented attacker variants. Because pre-defense recovery is effectively 100%, RR near 50 percentage points corresponds to post-sanitization bit accuracy at binary chance. Its main empirical advantage is stable near-chance sanitization with substantially smaller weight-distribution shift than the evaluated near-chance baselines PatternMask (PM) and Post-Training Quantization (PTQ).

Introduction

The rapid adoption of pretrained deep learning models has transformed computer vision, natural language processing, and cybersecurity (Vaswani et al. 2017; Devlin et al. 2019; Dosovitskiy et al. 2021). Repositories such as Hugging Face Model Hub now host hundreds of thousands of model checkpoints (Hugging Face 2026), enabling practitioners to deploy state-of-the-art Transformer architectures without retraining. However, this convenience introduces a new and underappreciated attack surface: the potential for model weights to act as covert carriers for malicious content (Wang, Liu, and Cui 2021; Liu et al. 2020; Hitaj et al. 2022). Unlike traditional

malware distribution, large-scale model files contain millions of 32-bit floating-point parameters, providing vast capacity for hiding malicious code with negligible impact on downstream accuracy. EvilModel (Wang, Liu, and Cui 2021) embeds up to 36.9 MB of malware in AlexNet through parameter substitution with minimal accuracy degradation. StegoNet uses neural-network parameters, including LSB substitution, as payload channels, whereas MaleficNet combines spread-spectrum embedding with LDPC error correction (Liu et al. 2020; Hitaj et al. 2022). Model X-Ray (Gilkarov and Dubin 2026) introduced detection frameworks using few-shot learning to identify payloads hidden through LSB substitution. Casey et al. (Casey, Santos, and Mirakhorli 2024) report widespread unsafe serialization and manually confirm 14 malicious Hugging Face files; a large-scale model-hub study likewise treats model distribution as a software supply-chain boundary (Zhao et al. 2024). We refer to these threats as *maleficent neural networks*: models that are active carriers of malware, not merely targets of adversarial attacks. Real-world incidents confirm the urgency. In 2022, a trojanized PyTorch dependency (`torchtriton`) was distributed via PyPI, a supply-chain vector structurally analogous to malicious model checkpoints (PyTorch Foundation 2022).

Despite growing recognition, existing defenses face fundamental limitations. Detection-oriented approaches identify suspicious checkpoints but do not remove the payload (Gilkarov and Dubin 2026). Post-training quantization reduces numerical precision and can overwrite low-order information as a side effect (Krishnamoorthi 2018; Nagel et al. 2020). Fine-tuning is effective but expensive and requires trusted data. Prior malware-in-weights work already includes error-correcting codes (Hitaj et al. 2022); the remaining gap is an explicit, scoped channel-erasure argument for complete replacement of a declared payload-bearing bit set. To address this gap, we propose GRAYSHIELD, a Gray-code-guided bit-level sanitizer that completely overwrites the target mantissa LSBs and uses a keyed tensor-specific phase for pattern diversity. We pose four research questions (RQs):

RQ1 (Threat Surface): At what LSB overwrite depth does mantissa-bit payload injection remain stealthy, lossless, and undetectable across diverse Transformer families?

RQ2 (Defense Effectiveness): Which post-training sanitiza-

^{*}Corresponding author.

tion strategies disrupt payload recovery to chance level while preserving downstream utility, under a Tier-1 naive attacker?

RQ3 (Adaptive Robustness): Does GRAYSHIELD maintain chance-level payload disruption when attackers strengthen their encoding strategy with redundancy coding and burst-resilient interleaving?

RQ4 (Deployment Trade-Off): Is GRAYSHIELD Pareto-optimal in the joint (Recovery Reduction, accuracy drop, variance, runtime) space, and what is the recommended deployment operating point?

Contributions. Our contributions are threefold: (i) we formalize a declared weight-LSB covert channel and show that complete payload-independent overwrite erases that target channel; (ii) we introduce a Gray-code overwrite structure with a keyed per-tensor phase for low-transition patterns and pattern diversity; and (iii) we evaluate security, utility, stability, fidelity, and runtime across four Transformer presets, two payloads, seven defenses, and five attacker variants, finding the smallest W_1 shift among the evaluated near-chance defenses.

Background and Related Work

Floating-point representation and LSB steganography. Deep learning frameworks store model parameters as IEEE-754 binary32 (FP32) values (Institute of Electrical and Electronics Engineers 2019). Each `float32` occupies 32 bits: one sign bit, eight biased exponent bits, and twenty-three mantissa bits. For a normalized FP32 value, changing the least significant fraction bit moves the value by one unit in the last place at its current exponent; the absolute perturbation therefore depends on that exponent. LSB steganography (Cox et al. 2007; Fridrich 2009) exploits this: an attacker reinterprets each weight as a 32-bit unsigned integer via bitcast, replaces the lowest x mantissa bits with payload bits, and writes back: $u'_i = (u_i \& \sim m_x) | (p_i \& m_x)$, $m_x = (1 \ll x) - 1$. Extraction is the exact inverse. StegoNet directly studies LSB substitution in neural-network parameters, and MaleficNet demonstrates coded weight payloads (Liu et al. 2020; Hitaj et al. 2022).

Supply chain and payload characteristics. Public repositories distribute pretrained checkpoints via a single API call without inspecting weight tensors. The difficulty of sanitization scales with payload entropy: low-entropy payloads (normalized entropy $\lesssim 0.65$) tolerate partial corruption; high-entropy payloads (entropy $\gtrsim 0.90$) require near-lossless reconstruction.

Defenses and their limitations. Detection-oriented methods (Gilkarov and Dubin 2026) identify compromised checkpoints but do not remove the payload. Task-aware PTQ methods optimize weight-rounding decisions using calibration data (Nagel et al. 2020); our evaluated static symmetric-INT8 quantize-dequantize baseline is a payload-disruption comparison, not a security-certified PTQ method. Active sanitization (Dubin 2023) directly perturbs weights, but perturbation-based approaches offer no complete-overwrite guarantee for the declared payload channel. Weight-bit integrity attacks and defenses instead study inference failure under sparse malicious flips (Yao, Rakin, and Fan 2020; He

et al. 2020); that objective differs from erasing an arbitrary byte stream stored in otherwise functional weights.

Gray codes. A Gray code (Gray 1953) is a bijection $\text{Gray} : \{0, \dots, 2^n - 1\} \rightarrow \{0, 1\}^n$ such that consecutive codewords differ in exactly one bit: $\text{Ham}(\text{Gray}(i), \text{Gray}(i+1)) = 1$. Standard reflected binary construction: $\text{Gray}(z) = z \oplus (z \gg 1)$. Two properties motivate its use here. (i) *Low transition*: adjacent unmasked codewords differ in one bit, inducing structured rather than independently scattered overwrite values. (ii) *Balance over complete masked periods*: for $g_i = \text{Gray}(i + \mathcal{O}_\ell) \& m_x$, the low x output bits are balanced over a period of 2^{x+1} ; the phase rotates the finite tensor segment and diversifies patterns. These properties motivate the design and its empirical fidelity; zero target-channel capacity follows from complete payload-independent overwrite, not from Gray coding or phase secrecy.

Related work on coding theory and steganography (Cox et al. 2007; Fridrich 2009; Shannon 1948) establishes that reliable communication requires a positive channel capacity. We are not aware of prior work that explicitly formulates complete replacement of a declared neural-weight payload channel as channel erasure. Model compression methods (Han, Mao, and Dally 2016; Frantar et al. 2023) are designed for efficiency, not security, and provide no guarantees against structured payload embedding.

GrayShield

Algorithm

Keyed phase for pattern diversity. For each target tensor $W^{(\ell)} \in \mathbb{R}^{N_\ell}$ (attention Q/K/V/O and MLP weight matrices in encoder blocks):

$$\mathcal{O}_\ell = H(\text{name}_\ell, x, \text{layer}_\ell, r; s) \bmod (2^{31} - 1), \quad (1)$$

where H is HMAC-SHA256 (Bellare, Canetti, and Krawczyk 1996), s is a 128-bit secret seed, and r is a reproducible run salt. The phase varies the overwrite pattern across tensors and releases. Secrecy is not required for target-channel erasure; it is useful only when pattern prediction or reuse is part of the deployment threat model.

Gray-code overwrite rule. For each element index i :

$$g_i^{(\ell)} = (n_i^{(\ell)} \oplus (n_i^{(\ell)} \gg 1)) \& m_x, \quad n_i^{(\ell)} = i + \mathcal{O}_\ell, \quad (2)$$

where $m_x = (1 \ll x) - 1$. The sanitized uint32 view is:

$$u'_i = (u_i \& \sim m_x) | g_i^{(\ell)}, \quad (3)$$

preserving sign and exponent bits, with adjacency smoothness $\text{Ham}(g_i^{(\ell)}, g_{i+1}^{(\ell)}) \leq 1$. Algorithm 1 gives the complete procedure. The operation is fully vectorizable, completing in $O(N)$ time with $O(1)$ auxiliary state per tensor.

Formal Guarantees

Proposition 1 (Target-Channel Erasure): Let P be any payload encoded only in a declared target-bit set $\mathcal{T}$, and let G be an overwrite string generated independently of P . If a sanitizer replaces every bit in $\mathcal{T}$ by G and the extractor reads only $\mathcal{T}$, then

$$\hat{P} = G, \quad \mathcal{I}(P; \hat{P}) = 0.$$

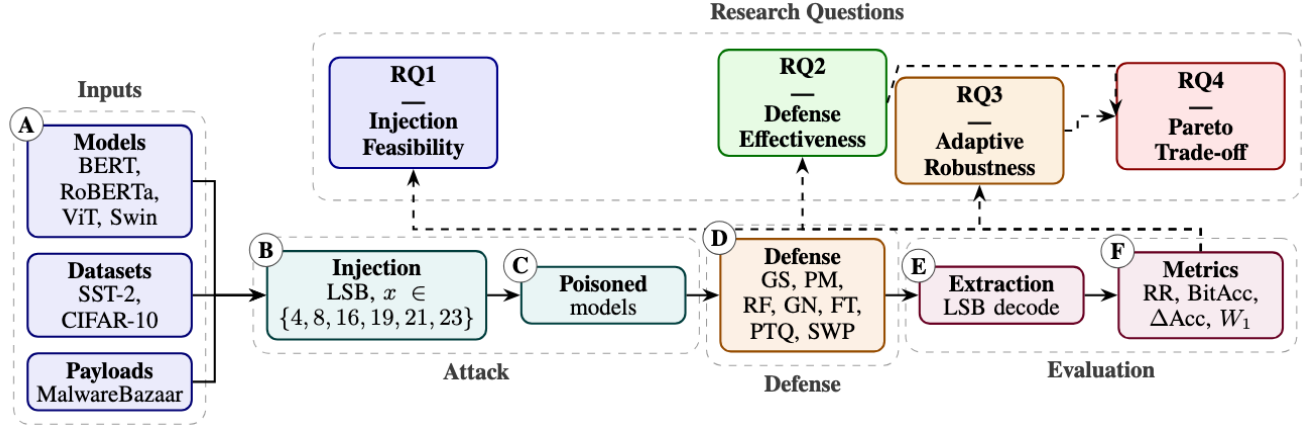


Figure 1: GrayShield evaluation pipeline. (A) Inputs define the empirical study space: pretrained models, evaluation datasets, and malware payloads from MalwareBazaar. (B) LSB injection overwrites the lowest x mantissa bits at depths $x \in \{4, 8, 16, 19, 21, 23\}$ to produce (C) poisoned models. (D) Seven sanitization strategies transform poisoned checkpoints into defended models. (E) Payload recovery is assessed by LSB extraction, and (F) security, fidelity, and utility metrics are computed. All four research questions are answered from the complete pipeline: RQ1 characterises the capacity boundary of injection; RQ2 compares defense effectiveness against the Tier-1 naive attacker; RQ3 evaluates adaptive robustness under five Tier-2 attacker variants (re-running stages B–F with varied attacker encodings); and RQ4 synthesises the Pareto trade-off from RQ2–RQ3 results. Dashed arrows denote analytical dependencies across research questions.

Algorithm 1 GRAYSHIELD Sanitization

Require: Transformer $\mathcal{M}$, depth x , secret seed s , salt r
Ensure: Sanitized $\mathcal{M}'$ with all declared target bits replaced

- 1: **for** each target tensor $W^{(\ell)}$ in $\mathcal{M}$ **do**
- 2: $\mathcal{O}_\ell \leftarrow H(\text{name}_\ell, x, \text{layer}_\ell, r; s) \bmod (2^{31} - 1)$
- 3: $u \leftarrow \text{bitcast}_{\text{uint32}}(W^{(\ell)})$
- 4: **for** each index i **do**
- 5: $n \leftarrow i + \mathcal{O}_\ell$
- 6: $g \leftarrow (n \oplus (n \gg 1)) \& m_x$
- 7: $u[i] \leftarrow (u[i] \& \sim m_x) | g$
- 8: **end for**
- 9: $W^{(\ell)'} \leftarrow \text{bitcast}_{\text{float32}}(u)$
- 10: **end for**
- 11: **return** $\mathcal{M}'$

Consequently the channel from P to the sanitized target bits has zero capacity for every payload distribution. This result requires neither Gray coding nor a secret seed.

Proof. Complete overwrite removes every target bit before writing G . Thus $p(\hat{P} = g \mid P = p) = p(G = g)$ for every p , so the joint distribution factorizes and $\mathcal{I}(P; \hat{P}) = 0$. Maximizing over input distributions still gives zero capacity. $\square$

Corollary 2 (Chance-Level Recovery): If, in addition, each overwritten bit is marginally balanced and independent of the corresponding payload bit, then $\mathbb{E}[\text{BitAcc}_{\text{post}}] = \frac{1}{2}$. When pre-defense recovery is effectively one, the empirical RR approaches 50 percentage points. Failing to reject $H_0^{(2c)}$ and $H_0^{(3c)}$ at $\alpha = 0.05$ constitutes empirical support for this corollary, while rejecting it for other defenses quantifies how

far those methods fall short of the channel-erasure bound.

Because the sanitized target bits are independent of P , data processing prevents any decoder from reconstructing information absent from the extractor input. Under the balanced condition in Corollary 5, the bitwise view is BSC($\frac{1}{2}$) and Shannon’s converse excludes reliable positive-rate coding (Shannon 1948), including Reed–Solomon, LDPC, and polar families (Reed and Solomon 1960; Gallager 1962; Arikan 2009).

Theorem 3 (Conditional Loss Bound): If $\mathcal{L} : \mathcal{W} \rightarrow \mathbb{R}$ is L_{lip} -Lipschitz in the model parameters, then

$$\begin{aligned} |\mathcal{L}(W') - \mathcal{L}(W)| &\leq L_{\text{lip}} \|W' - W\|_2, \\ &= L_{\text{lip}} \|W\|_2 D_{\text{rel}}(W, W'). \end{aligned}$$

In the completed payload set, worst-case $D_{\text{rel}} = 7.1 \times 10^{-4}$ at $x = 16$ and 5.4×10^{-3} at $x = 19$. This theorem bounds loss only; accuracy is evaluated empirically. Further assumptions, scope details, and reduced-precision analysis are provided in the supplementary material.

GRAYSHIELD versus PatternMask. Both GRAYSHIELD and a fixed full-overwrite PatternMask satisfy Proposition 4 for the declared target channel. Their distinction is therefore empirical and operational, not target-channel capacity: GRAYSHIELD uses a low-transition Gray structure and a keyed phase for pattern diversity, whereas PatternMask reuses one global pattern. In the evaluated configurations, GRAYSHIELD has substantially lower RR variance and W_1 shift than PatternMask despite similar mean RR. Erasure follows from payload-independent overwriting, while the keyed phase provides pattern diversity and the Gray-code structure preserves distributional fidelity; phase secrecy is not required for erasure.

Threat Model

We consider a supply-chain adversary who embeds a binary payload into the mantissa LSBs of FP32 weights of a pre-trained Transformer distributed through a public repository. The adversary has full white-box access to the model prior to distribution; the defender applies post-training sanitization before publication and has no access to the training data.

Target parameter set and hiding capacity We target the *encoder blocks* of the Transformer backbone: attention (Q/K/V/O) and feed-forward (MLP) weight matrices, which contain the dominant fraction of trainable parameters. Embedding layers and task-specific classifier heads are excluded by default. The hiding capacity at overwrite depth x is:

$$C(\theta_{\text{target}}, x) = \sum_{w \in \theta_{\text{target}}} \text{numel}(w) \cdot x, \quad x \in [1, 23]. \quad (4)$$

Three-tier attacker knowledge hierarchy

- **Tier 1 (Black-box).** Knows architecture and parameter count; uses standard LSB substitution without knowledge of the defender’s method. Corresponds to NAIVE.
- **Tier 2 (Gray-box).** Knows a sanitization step of unspecified type will be applied; responds with redundancy coding (REPEAT3/REPEAT5) or burst-resilient interleaving.
- **Tier 3 (White-box).** Knows the sanitization algorithm (GRAYSHIELD) but not the seed s . Strongest realistic attacker.

Experimental Setup

Models and datasets. We evaluate across both natural language and vision domains using standard pretrained Transformer architectures. For text, we use BERT-base (Devlin et al. 2019) and RoBERTa-base (Liu et al. 2019) fine-tuned on the SST-2 sentiment classification benchmark (Socher et al. 2013; Wang et al. 2018). For vision, we use ViT-Base/16 (Dosovitskiy et al. 2021) and Swin-Base (Liu et al. 2021) evaluated on CIFAR-10 (Krizhevsky 2009). These models span a range of architectural designs and parameter scales, providing coverage across both language and image modalities. All evaluations are conducted on held-out subsets of size $n_{\text{eval}} = 512$, yielding an accuracy resolution of approximately 0.195 pp, which is sufficient to reliably detect sub-1% performance changes.

Malware payloads. Two representative Malware-Bazaar payloads (abuse.ch 2026) spanning the entropy extremes: a low-entropy (LE) JavaScript script (c37...a88.js.malware, 134.3 KB, entropy 0.603) and a high-entropy (HE) compiled binary (5704...6d0.exe.malware, 8,971.4 KB, entropy 0.994). The LE payload tolerates partial corruption due to structural redundancy; the HE payload requires near-lossless reconstruction.

Defenses. Seven post-training strategies evaluated at operating depth $x = 19$: GS (proposed), PM (fixed-pattern LSB overwrite), PTQ (symmetric INT8 quantize-dequantize (Krishnamoorthi 2018)), FT (fine-tune on 256 clean samples, sweeping $\{50, 100, 200, 500\}$ gradient steps), GN (i.i.d. $\mathcal{N}(0, \sigma^2)$ noise, sweeping $\sigma \in$

$\{10^{-6}, 3 \times 10^{-6}, 10^{-5}, 3 \times 10^{-5}, 10^{-4}\}$), SWP (an author-defined, magnitude-ranked selective-weight perturbation baseline), and RF (Bernoulli bit flip). We set the defense depth to the largest utility-preserving value ($x = 19$, RQ1), covering all stealthy embeddings ($x' \leq x$); deeper embeddings ($x' > x$) exhibit detectable accuracy and geometric degradation. **Attacker variants.** Five Tier-2 (gray-box) variants: NAIVE, INTERLEAVE, REPEAT3, REPEAT5, and RS(255,127). The threat model is detailed in the supplementary material.

Recovery metric. For payload bits $p_1, \dots, p_B$ and extracted bits $\hat{p}_1, \dots, \hat{p}_B$,

$$\text{Rec} = B^{-1} \sum_{i=1}^B \mathbf{1}[p_i = \hat{p}_i]; \quad \text{RR} = 100(\text{Rec}_{\text{pre}} - \text{Rec}_{\text{post}}).$$

We report RR in percentage points (pp), not as a relative percentage. Pre-defense recovery is effectively one in the completed runs; hence RR near 50 pp means post-defense bit accuracy near binary chance. Values farther above 50 pp are not automatically more secure because systematic anti-correlation can be inverted. **Fidelity and cost metrics.** Beyond RR and ΔAcc , we report weight-distribution fidelity (W_1), runtime (ms), RR consistency (fraction within ± 2 pp of the mean), and bit-level Hamming distance (millions of bits). **Statistical protocol.** All experiments use run seed 1, evaluation seed 42. The supplementary material pre-specifies 13 null hypotheses; all tests two-tailed at $\alpha = 0.05$ unless marked one-tailed ($\uparrow$). Effect sizes reported alongside p -values: η_{KW}^2 for Kruskal-Wallis, Cohen’s d for t -tests, Spearman ρ for correlations, Kendall’s W for Friedman, Wilks’ λ for MANOVA.

Results

RQ1: Injection Feasibility

RQ1 asks whether mantissa-LSB payload injection can remain stealthy while preserving model utility. Fig. 2 provides the consolidated evidence across six overwrite depths $x \in \{4, 8, 16, 19, 21, 23\}$, four Transformer model presets, and two entropy-stratified payloads. Two patterns are central. First, both geometric perturbation metrics increase monotonically with overwrite depth. Cosine deviation and relative ℓ_2 remain negligible through $x = 16$, but grow rapidly beyond that point. Panels (a) and (b) report perfect monotone association across the six depth-wise means (Spearman $\rho = 1.000$, $p < 0.001$), identifying overwrite depth as the primary driver of geometric detectability.

Second, the utility boundary is sharper for high-entropy payloads. In Fig. 2(c), low-entropy payloads remain essentially flat across all tested depths, whereas high-entropy payloads begin to show measurable degradation at $x \geq 19$. Even then, the worst completed case remains below the 1% threshold. The grouped Mann-Whitney result ($p = 0.005$) supports the conclusion that the accuracy impact is significantly larger above the $x = 16$ boundary. $H_0^{(1b)}$: **Reject**. $H_0^{(1c)}$: **Reject**.

Answer to RQ1. Payload injection remains stealthy through $x = 16$ and enters a transition regime at $x = 19$. We treat $x = 19$ as the main operating point for RQ2–RQ4:

it represents a high-capacity attacker while remaining below the practical utility-loss threshold.

RQ2: Defense Effectiveness under a Naive Attacker

RQ2 compares seven post-training defenses at the RQ1-selected operating point. The goal is not merely to approach the 50 pp chance target, but to identify a defense that simultaneously disrupts payload recovery, preserves accuracy, avoids excessive distribution shift, and remains operationally lightweight. Table 1 and Fig. 3 show that GS, PM, and PTQ are the only defenses that reach chance-level sanitization under the naive attacker. Their mean RR values differ by less than one percentage point. However, they are not equivalent once secondary deployment criteria are considered.

PM attains mean RR of 50.00 pp but with substantially higher variance (± 2.64) than GS (± 0.54). PTQ is similarly effective in RR (49.93 pp) but requires calibration data and induces a larger distributional shift. GN is unstable, with RR 31.49 ± 15.08 pp, while SWP and RF are ineffective in this setting.

$H_0^{(2a)}$: **Reject** ($H = 48.2$, $p < 0.001$, $\eta_{KW}^2 = 0.861$).

$H_0^{(2b)}$: **Reject** ($H = 43.1$, $p < 0.001$, $\eta_{KW}^2 = 0.803$).

$H_0^{(2c)}$: **Fail to reject** for GS, PM, PTQ ($p \in [0.21, 0.99]$, $|d| \leq 0.52$); **Reject** for FT, GN, SWP, RF ($p \leq 0.046$).

$H_0^{(2d)}$: **Reject** ($W = 7.53$, $p = 0.016$, $\sigma_{GN}/\sigma_{GS} = 27.9\times$).

Fig. 3(a)–(c) clarifies that mean RR alone is insufficient. GS provides the most balanced profile: near-chance recovery reduction, negligible impact on accuracy, zero data dependency, low variance, and low distribution shift. GS achieves average $W_1 = 2.6 \times 10^{-5}$, approximately $25\times$ smaller than PM’s 7.1×10^{-4} and $42\times$ smaller than PTQ’s 1.17×10^{-3} .

Answer to RQ2. Under the naive attacker, GS, PM, and PTQ all disrupt payload recovery to near chance level. GS is the most deployable among them because it combines near-chance RR with low variance, no data requirement, small utility loss, and the smallest W_1 shift in this high-RR group. Both exact full-overwrite methods, GS and PM, satisfy the scoped target-erasure result; their empirical stability and fidelity differ.

RQ3: Robustness under Adaptive Attackers

RQ3 evaluates whether the RQ2 ranking survives stronger attacker strategies. The five implemented variants are NAIVE, INTERLEAVE, REPEAT3, REPEAT5, and RS(255,127). Table 2 reports RR per attacker variant. The defining result is GS’s invariance to attacker adaptation. Across all five attacker variants, GS stays within 49.5–50.4 pp, with aggregate RR 49.96 ± 0.66 pp. The RQ2 means in Table 1 and the NAIVE column in Table 2 use different run pools: RQ2 contains all eight completed model–payload pairs, whereas RQ3 contains four completed pairs and omits the zero-noise GN control. Their Naive values therefore need not coincide. This agrees with target-channel erasure: coding cannot reconstruct payload information absent from the extractor input. PM remains near 50 pp but has much larger dispersion ($\sigma = 3.25$ pp); PTQ is slower and more distortive, while GN, SWP, and RF degrade under adaptive variants. The one-sample compari-

son printed in Fig. 4(c) fails to distinguish the five attacker-specific GS means from the 50 pp target ($p = 0.822$).

Fig. 4 makes this difference visually explicit. Panel (a) shows GS as nearly flat across attackers; panel (b) shows a compact high-RR distribution; panel (c) places GS in the desired effectiveness–stability region; and panel (d) confirms that GS does not exhibit a strategy-specific weakness under interleaving, repetition, or Reed–Solomon coding.

Answer to RQ3. GS maintains stable chance-level sanitization under all five implemented adaptive attacker variants.

RQ4: Pareto Trade-off and Deployment Profile

RQ4 integrates the previous findings into a deployment-oriented comparison. A practical defense must satisfy more than high mean RR: it must also preserve accuracy, remain stable, avoid excessive runtime, and limit distributional drift.

Figure 5(a) shows the RR–accuracy Pareto plane. GS lies in the desired region: high RR, low $|\Delta Acc|$, and low variance. PM has similar mean RR and low runtime but fails the stability criterion ($\sigma = 3.25$). PTQ has high RR and low variance but fails the runtime criterion (882 ms) and has the largest W_1 among the high-RR defenses. FT, GN, SWP, and RF are dominated by at least one of the high-RR methods. Figure 5(b) summarizes the deployment constraints. GS is the only defense satisfying all four criteria simultaneously: RR in $[48, 52]$ pp, $\sigma_{RR} < 2$ pp, $|\Delta Acc| < 1\%$, and runtime < 500 ms. PM and PTQ satisfy three of four, but each misses a different requirement.

In Fig. 5(c), GS requires 54 ms per checkpoint, compared with 882 ms for PTQ and 13.2 s for FT. In Fig. 5(d), PM’s W_1 is approximately $25\times$ larger than GS’s, and PTQ’s approximately $42\times$ larger.

Answer to RQ4. GS at $x = 19$ is the recommended operating point. It achieves 49.96 ± 0.66 pp RR, sub-1% accuracy impact, low variance, low runtime, and substantially better weight-distribution fidelity than competing high-RR baselines.

Discussion

Theory–empirics alignment. Complete payload-independent replacement erases the declared target channel without seed secrecy; balance is the additional condition that makes bit accuracy chance-level. The reported 49.96 ± 0.66 pp RR across five attackers is consistent with both properties: repetition, interleaving, and RS coding cannot reconstruct information absent from the extractor input. **Stability and fidelity.** GS and PM share the scoped full-overwrite guarantee and similar mean RR, but PM has higher dispersion and $25\times$ larger W_1 . GN’s lower and unstable RR despite matched perturbation magnitude further shows that targeting the payload-bearing LSB subspace matters more than perturbation magnitude alone.

Deployment interpretation. Table 2 and Supplementary Fig. S1 show that GS alone satisfies the complete operating envelope: RR in $[48, 52]$ pp, $\sigma_{RR} < 2$ pp, $|\Delta Acc| < 1\%$, runtime below 500 ms, and low distributional drift. PTQ is slower and more distortive, PM is less stable, and FT is slow and requires trusted data.

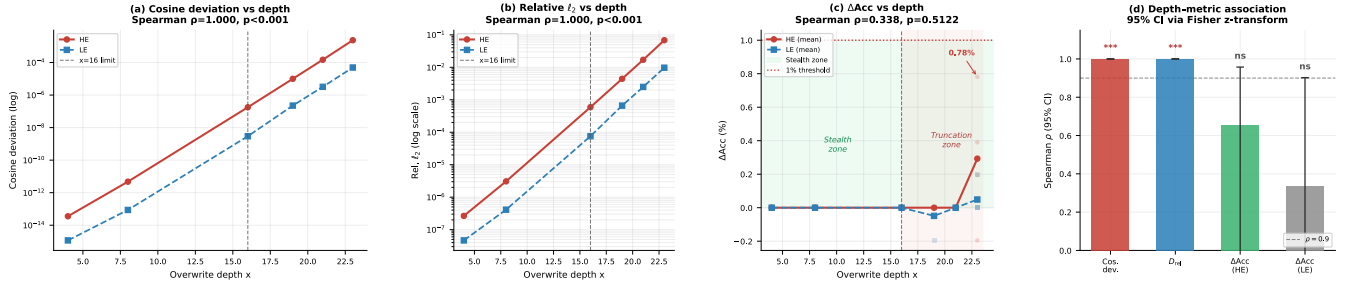


Figure 2: RQ1 — Injection feasibility across six overwrite depths ($x \in \{4, 8, 16, 19, 21, 23\}$), four Transformer architectures, and two entropy-stratified payloads ($n = 48$ runs). (a) Cosine deviation vs. depth on a log scale; Spearman $\rho = 0.954$, $p < 0.001$. (b) Relative ℓ_2 distance vs. depth on a log scale; $\rho = 0.954$, $p < 0.001$. (c) Signed accuracy drop ΔAcc vs. depth. The green region ($x \leq 16$) marks depths satisfying the stealth constraints, while the red region ($x \geq 19$) marks the onset of measurable degradation, with a worst-case high-entropy drop of 0.78%. Mann–Whitney comparison between low and high depth regimes yields $p = 0.005$. (d) Spearman depth–metric association with 95% Fisher- z confidence intervals.

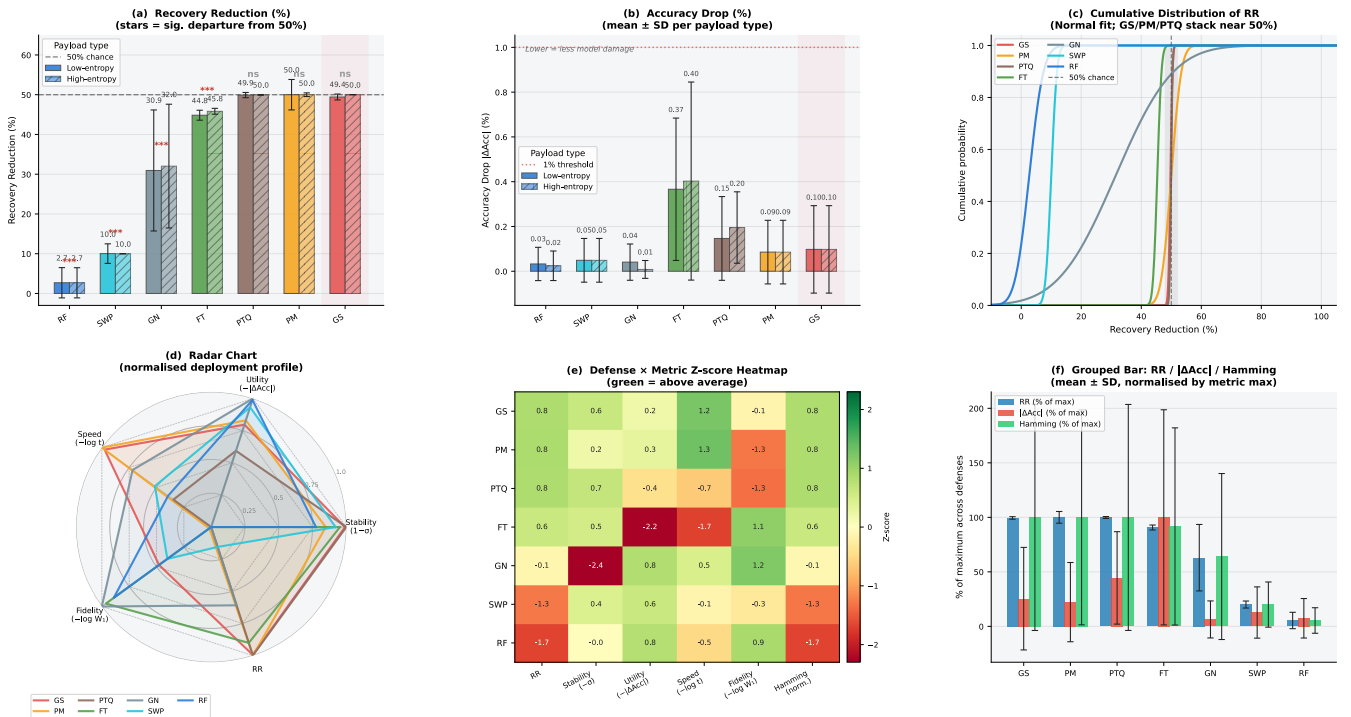


Figure 3: RQ2 — Defense effectiveness and multi-metric trade-offs. (a) Recovery Reduction (RR) across defenses, stratified by payload entropy. GrayShield reaches near-chance disruption (≈ 50 pp) consistently. (b) Accuracy drop ($|\Delta\text{Acc}|$), where lower values indicate better utility preservation. (c) Cumulative distribution of RR, showing that GS, PM, and PTQ concentrate near the 50 pp target while weaker defenses remain far below. (d) Radar chart of normalized deployment metrics. (e) Defense-by-metric Z-score heatmap, where green indicates above-average performance. (f) Grouped normalized bar chart.

Limitations. Experiments cover FP32 and four 86M–125M presets; reduced-precision utility remains unvalidated. LDPC and polar attackers are analytical extensions, and some FT RQ3 runs were incomplete. The guarantee is bit-local: higher-bit payloads, executable serialization, metadata side channels, and post-sanitization reinsertion are outside scope. Executable malware samples are not redistributed.

Generative-AI assistance. Generative-AI tools assisted with language editing, $\text{\LaTeX}$ formatting, and verification; the authors reviewed the output and retain full responsibility.

Conclusion

GRAYSHIELD completely replaces a declared weight-LSB payload channel using a low-transition Gray sequence. Payload-independent overwrite gives the scoped zero-

Defense	Data	RR (pp)	ΔAcc	W_1	Target erasure
GS	None	49.72 ± 0.54	-0.10%	2.6×10^{-5}	Prop. 4
PM	None	50.00 ± 2.64	$+0.09\%$	7.6×10^{-4}	Prop. 4
PTQ	Calibration	49.93 ± 0.40	$+0.07\%$	8.5×10^{-4}	None
FT	Clean set	45.34 ± 1.11	$+0.38\%$	7.6×10^{-7}	None
GN	None	31.49 ± 15.08	-0.02%	5.8×10^{-7}	None
SWP	None	10.01 ± 1.50	$+0.05\%$	4.6×10^{-5}	None
RF	None	2.69 ± 3.72	$+0.01\%$	1.3×10^{-6}	None

Table 1: RQ2 defense comparison. RR denotes Recovery Reduction (pp); values near 50 pp indicate chance-level recovery, the target under complete payload-independent overwrite.

D	Naïve	Interl.	Rep3	Rep5	RS	RR	σ	ΔAcc	Time	Cons.	Ham.	W_1
GS	49.5 ± 0.7	50.0 ± 0.2	50.4 ± 0.6	50.2 ± 0.2	49.6 ± 0.8	49.96	0.66	-0.10	54	0.99	9.60	2.8×10^{-5}
PM	50.0 ± 3.2	50.0 ± 3.2	50.0 ± 3.3	50.0 ± 3.3	50.0 ± 3.2	50.00	3.25	$+0.13$	44	0.94	9.60	7.1×10^{-4}
PTQ	50.0 ± 0.5	50.0 ± 0.5	49.5 ± 0.5	50.1 ± 0.5	49.5 ± 0.4	49.83	0.54	$+0.17$	882	0.99	9.57	1.2×10^{-3}
FT	45.2 ± 1.4	45.3 ± 1.3	45.5 ± 1.0	46.3 ± 0.9	45.2 ± 1.1	45.51	1.23	$+0.45$	13,200	0.97	8.93	4.0×10^{-6}
GN	37.5 ± 6.1	37.7 ± 6.0	38.4 ± 5.6	40.3 ± 4.9	37.5 ± 6.0	38.25	5.85	-0.04	620	0.85	7.71	4.0×10^{-6}
SWP	10.0 ± 2.1	10.1 ± 2.1	9.5 ± 0.9	8.6 ± 0.9	5.5 ± 2.7	8.76	2.56	$+0.07$	1,480	0.71	1.67	4.7×10^{-5}
RF	2.7 ± 3.7	2.7 ± 3.7	0.6 ± 1.0	0.2 ± 0.3	2.5 ± 3.8	1.73	3.16	$+0.02$	2,590	0.00	0.33	4.0×10^{-6}

Table 2: Comprehensive defense (D) evaluation across RQ2–RQ3 ($x = 19$). Attacker columns report mean RR (pp) $\pm$ std. Summary reports mean RR, std, ΔAcc (%), runtime (ms), and consistency. Hamming distance is reported in millions (M).

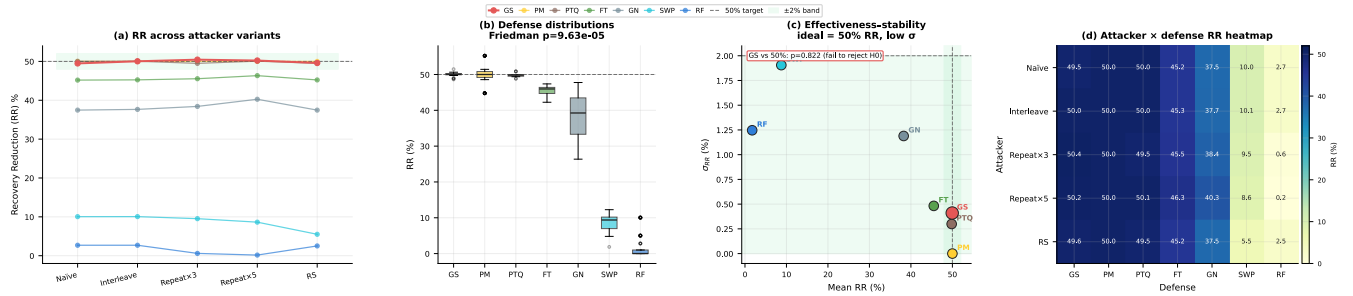


Figure 4: RQ3 — Adaptive attacker robustness. (a) RR across attacker variants shows that GS remains stable around 50 pp. (b) Defense-level RR distributions show significant differences across defenses, with GS maintaining high and consistent performance. (c) Effectiveness–stability trade-off shows GS closest to the target region of high RR and low σ . (d) Attacker-by-defense heatmap confirms that GS remains resilient across all implemented adaptive variants.

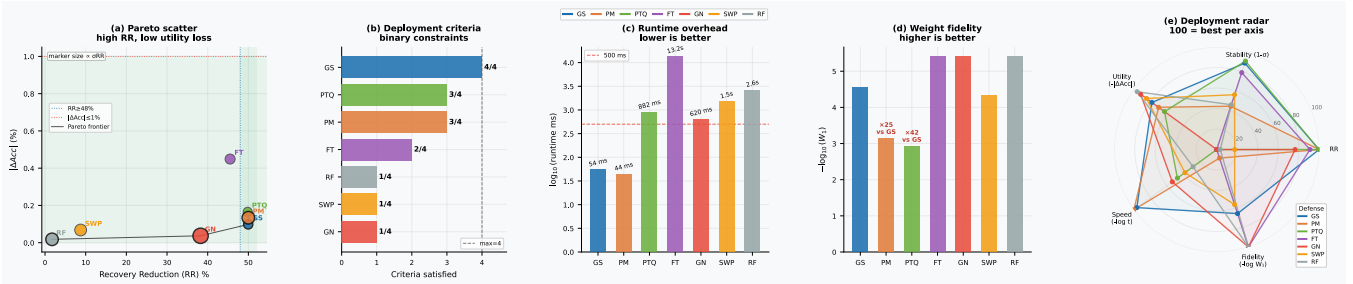


Figure 5: RQ4 deployment profile: (a) RR–accuracy Pareto plane; (b) criteria satisfied; (c) runtime; (d) $-\log_{10}(W_1)$ fidelity; and (e) normalized deployment radar.

capacity result; Gray structure and the keyed phase provide regularity and pattern diversity. Across the evaluated settings, GRAYSHIELD achieves stable near-chance recovery with sub-

1% accuracy degradation, and combines low variance and runtime with lower W_1 than the evaluated near-chance base-lines.

References

- abuse.ch. 2026. MalwareBazaar: Malware Sample Exchange. <https://bazaar.abuse.ch/>. Accessed: 2026-07-24.
- Anonymous Authors. 2026. GrayShield: Anonymized Replication Package. Anonymized replication package for double-blind review. Contains GrayShield sanitization code, attacker implementations, configuration grids, figure- and table-generation scripts, curated payload data, dataset metadata, structured result artifacts, and provenance information.
- Arikan, E. 2009. Channel Polarization: A Method for Constructing Capacity-Achieving Codes for Symmetric Binary-Input Memoryless Channels. *IEEE Transactions on Information Theory*, 55(7): 3051–3073.
- Bellare, M.; Canetti, R.; and Krawczyk, H. 1996. Keying Hash Functions for Message Authentication. In *Advances in Cryptology—CRYPTO '96*, volume 1109 of *Lecture Notes in Computer Science*, 1–15.
- Casey, B.; Santos, J. C. S.; and Mirakhorli, M. 2024. A Large-Scale Exploit Instrumentation Study of AI/ML Supply Chain Attacks in Hugging Face Models. *arXiv:2410.04490*.
- Cox, I. J.; Miller, M. L.; Bloom, J. A.; Fridrich, J.; and Kalker, T. 2007. *Digital Watermarking and Steganography*. Morgan Kaufmann, 2nd edition. ISBN 978-0-12-372585-1.
- Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 4171–4186.
- Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; Uszkoreit, J.; and Houlsby, N. 2021. An Image Is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *International Conference on Learning Representations*.
- Dubin, R. 2023. Disarming Attacks Inside Neural Network Models. *IEEE Access*, 11: 124295–124303.
- Frantar, E.; Ashkboos, S.; Hoefler, T.; and Alistarh, D. 2023. GPTQ: Accurate Post-Training Quantization for Generative Pre-trained Transformers. In *International Conference on Learning Representations*.
- Fridrich, J. 2009. *Steganography in Digital Media: Principles, Algorithms, and Applications*. Cambridge University Press. ISBN 978-0-521-19019-0.
- Gallager, R. G. 1962. Low-Density Parity-Check Codes. *IRE Transactions on Information Theory*, 8(1): 21–28.
- Gilkarov, D.; and Dubin, R. 2026. Model X-Ray: Detection of Hidden Malware in AI Model Weights Using Few-Shot Learning. *Journal of Information Security and Applications*, 101: 104517.
- Gray, F. 1953. Pulse Code Communication. US Patent 2,632,058.
- Han, S.; Mao, H.; and Dally, W. J. 2016. Deep Compression: Compressing Deep Neural Networks with Pruning, Trained Quantization and Huffman Coding. In *International Conference on Learning Representations*.
- He, Z.; Rakin, A. S.; Li, J.; Chakrabarti, C.; and Fan, D. 2020. Defending and Harnessing the Bit-Flip Based Adversarial Weight Attack. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14083–14091.
- Hitaj, D.; Pagnotta, G.; Hitaj, B.; Mancini, L. V.; and Perez-Cruz, F. 2022. MaleficNet: Hiding Malware into Deep Neural Networks Using Spread-Spectrum Channel Coding. In *Computer Security—ESORICS 2022*, volume 13556 of *Lecture Notes in Computer Science*, 425–444. Springer.
- Hugging Face. 2026. Hugging Face Model Hub. <https://huggingface.co/models>. Accessed: 2026-07-24.
- Institute of Electrical and Electronics Engineers. 2019. IEEE Standard for Floating-Point Arithmetic (IEEE Std 754-2019). IEEE.
- Krishnamoorthi, R. 2018. Quantizing Deep Convolutional Networks for Efficient Inference: A Whitepaper. *arXiv preprint arXiv:1806.08342*.
- Krizhevsky, A. 2009. Learning Multiple Layers of Features from Tiny Images. Technical report, University of Toronto.
- Liu, T.; Liu, Z.; Liu, Q.; Wen, W.; Xu, W.; and Li, M. 2020. StegoNet: Turn Deep Neural Network into a Stegomalware. In *Proceedings of the 36th Annual Computer Security Applications Conference*, 928–938. Association for Computing Machinery.
- Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; and Stoyanov, V. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv preprint arXiv:1907.11692*.
- Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; and Guo, B. 2021. Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 10012–10022.
- Nagel, M.; Amjad, R. A.; van Baalen, M.; Louizos, C.; and Blankevoort, T. 2020. Up or Down? Adaptive Rounding for Post-Training Quantization. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, 7197–7206. PMLR.
- PyTorch Foundation. 2022. Compromised PyTorch-Nightly Dependency Chain Between December 25th and December 30th, 2022. <https://pytorch.org/blog/compromised-nightly-dependency/>. Accessed: 2026-07-24.
- Reed, I. S.; and Solomon, G. 1960. Polynomial Codes over Certain Finite Fields. *Journal of the Society for Industrial and Applied Mathematics*, 8(2): 300–304.
- Sakai, M.; and Imaizumi, M. 2026. Spectrum-Adaptive Generalization Bounds for Trained Deep Transformers. *arXiv preprint arXiv:2605.07297*.
- Shannon, C. E. 1948. A Mathematical Theory of Communication. *Bell System Technical Journal*, 27(3): 379–423.
- Socher, R.; Perelygin, A.; Wu, J.; Chuang, J.; Manning, C. D.; Ng, A. Y.; and Potts, C. 2013. Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, 1631–1642.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention Is All You Need. In *Advances in Neural Information Processing Systems*, volume 30, 5998–6008.

Wang, A.; Singh, A.; Michael, J.; Hill, F.; Levy, O.; and Bowman, S. R. 2018. GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, 353–355.

Wang, Z.; Liu, C.; and Cui, X. 2021. EvilModel: Hiding Malware Inside of Neural Network Models. In *2021 IEEE Symposium on Computers and Communications*, 1–7.

Yao, F.; Rakin, A. S.; and Fan, D. 2020. DeepHammer: Depleting the Intelligence of Deep Neural Networks through Targeted Chain of Bit Flips. In *29th USENIX Security Symposium*, 1463–1480. USENIX Association.

Zhao, J.; Wang, S.; Zhao, Y.; Hou, X.; Wang, K.; Gao, P.; Zhang, Y.; Wei, C.; and Wang, H. 2024. Models Are Codes: Towards Measuring Malicious Code Poisoning Attacks on Pre-trained Model Hubs. In *Proceedings of the 39th IEEE/ACM International Conference on Automated Software Engineering*, 2087–2098. Association for Computing Machinery.

Supplementary Material

Pre-specified Hypotheses and Outcomes

Table 3 consolidates all 13 pre-specified null hypotheses for RQ1–RQ4, their tests, effect sizes, and decisions (R = Reject; FTR = Fail to Reject).

Table 3: Consolidated pre-specified statistical hypotheses for RQ1–RQ4. All tests are two-tailed at $\alpha = 0.05$ unless marked ($\uparrow$) one-tailed. *Bonferroni-corrected over $H_0^{(4b)} - H_0^{(4c)}$.

ID	Null hypothesis H_0	Test	Effect	Decision
<i>RQ1 — Injection Feasibility</i>				
$H_0^{(1a)}$	$\rho(\text{depth, metric}) = 0$	Spearman	$\rho = 0.954, p < 0.001$	R
	(LE ΔAcc)	Spearman	$\rho = 0.338, p = 0.512$	FTR
$H_0^{(1b)}$	Equal ΔAcc across 6 depths	KW ($n = 8$)	$\eta_{KW}^2, p < 0.001$	R
$H_0^{(1c)\uparrow}$	$\Delta\text{Acc}(x \geq 19) \leq \Delta\text{Acc}(x \leq 16)$	MW-U	$p = 0.005$	R
<i>RQ2 — Baseline Comparison (naive)</i>				
$H_0^{(2a)}$	Equal means RR across 7 defenses	KW ($n = 8$)	$\eta^2 = 0.861, p < 0.001$	R
$H_0^{(2b)}$	Equal $ \Delta\text{Acc} $ across 7 defenses	KW ($n = 8$)	$\eta^2 = 0.803, p < 0.001$	R
$H_0^{(2c)}$	$\mu_{RR}^{(m)} = 50\%$ (GS/PM/PTQ)	1-sample t	$ d \leq 0.52, p \geq 0.207$	FTR
	$\mu_{RR}^{(m)} = 50\%$ (FT/GN/SWP/RF)	1-sample t	$ d \geq 0.86, p \leq 0.046$	R
$H_0^{(2d)}$	$\sigma_{GS}^2 = \sigma_{GN}^2$	Levene ($n = 8$)	ratio $27.9\times, p = 0.016$	R
<i>RQ3 — Adaptive Robustness</i>				
$H_0^{(3a)}$	Equal mean RR (adaptive pool)	KW ($n = 5$)	$\eta^2 = 0.909, p < 0.001$	R
$H_0^{(3b)}$	Attacker does not reshuffle ordering	Friedman	$W = 0.81, p < 0.001$	R
$H_0^{(3c)*}$	$\mu_{RR}^{(GS)} = 50\%$ (adaptive)	1-sample t ($n = 5$)	$d = -0.16, p = 0.745$	FTR
$H_0^{(3d)}$	$\sigma_{GS}^2 = \sigma_{PM}^2$ (adaptive)	Levene+Bartlett	ratio $24.3\times, p = 0.010$	R
<i>RQ4 — Pareto Trade-off</i>				
$H_0^{(4a)}$	Equal multivariate profiles	MANOVA	$\lambda \approx 0, p < 0.001$	R
$H_0^{(4b)}$	Equal $ \Delta\text{Acc} $: GS vs PM	Welch t ($n = 8$)	$d = 0.42, p = 0.009^*$	R
$H_0^{(4c)}$	Equal $ \Delta\text{Acc} $: GS vs PTQ	Welch t ($n = 8$)	$d = 0.78, p < 0.001^*$	R

* **Key result.** $H_0^{(3c)}$ is the most critical single test in the paper. Failing to reject confirms that GRAYSHIELD induces a $\text{BSC}(\frac{1}{2})$ channel regardless of the coding strategy employed, empirically validating Proposition 4.

Formal Guarantees

Assumptions

Our theoretical analysis relies on the following standard assumptions:

- **PRF security.** HMAC-SHA256 is computationally indistinguishable from a uniformly random function by any PPT adversary (Definition 14).
- **Secret seed.** The sanitization seed s ($|s| = 128$ bits) is unknown to the attacker and stored securely (e.g., HSM).
- **Computational boundedness.** The attacker is limited to PPT algorithms.
- **Memoryless channel.** Post-sanitization extracted bits are conditionally independent across positions (Definition 20).

Information-theoretic framing

We model LSB-based payload embedding as a sequence of channel uses over a memoryless binary channel. Let $p_i \in \{0, 1\}$ be the i -th payload bit written by the attacker and $\hat{p}_i \in \{0, 1\}$ the bit recovered after sanitization T . Pre-defense, the channel is noiseless: $\hat{p}_i = p_i$ with probability one, giving raw capacity $C_{\text{raw}} = \sum_w \text{numel}(w) \cdot x$. The defender’s objective is to choose T such that:

$$C_{\text{eff}}(T) = \mathcal{I}(P; T(\text{lsb}_x(W))) \rightarrow 0. \quad (5)$$

Perfect Channel Erasure

Proposition 4 (Perfect Channel Erasure): Let $T_s : \mathcal{W} \rightarrow \mathcal{W}$ be the GRAYSHIELD sanitization map parameterized by secret seed s , and let $\hat{P} = \mathcal{D}(T_s(\mathcal{E}(W, P)))$ be the extracted payload after sanitization. Under the assumptions above, for any payload distribution $\mathcal{P}$:

$$\mathcal{I}(P; \hat{P}) = 0,$$

and therefore $C_{\text{eff}}(T_s) = 0$.

Proof. For each index i in a target tensor, the overwritten mantissa bits are replaced by $g_i = f_s(i) \in \{0, 1\}^x$, where $f_s(\cdot)$ is derived from a PRF keyed by s via the per-tensor phase offset $\mathcal{O}_\ell$ and the Gray-code mapping. By PRF security, f_s is computationally indistinguishable from a uniformly random function; consequently:

$$\Pr(g_{i,k} = 1) = \frac{1}{2} \quad \text{for each bit } k \in [0, x),$$

and $g_{i,k}$ is independent of the payload bit $P_{i,k}$ when s is hidden from the attacker. This induces a BSC($\frac{1}{2}$) on each transmitted bit, whose capacity is:

$$C_{\text{BSC}}(\frac{1}{2}) = 1 - H_b(\frac{1}{2}) = 0.$$

By the memoryless channel property, the extracted bits $\hat{P}_i$ are conditionally independent across positions, so:

$$\mathcal{I}(P; \hat{P}) = \sum_{i=1}^B \mathcal{I}(P_i; \hat{P}_i) = 0. \quad \square$$

Corollary 5 (Chance-Level Recovery): Under Proposition 4, $\mathbb{E}[\text{BitAcc}_{\text{post}}] = \frac{1}{2}$, so the empirical Recovery Reduction RR approaches 50%.

Coding-Theoretic Limits

Lemma 6 (Irrecoverability under Error-Correcting Codes): Let the sanitized channel be BSC($\frac{1}{2}$). Then for any error-correcting code with positive rate $R > 0$:

$$\lim_{n \rightarrow \infty} \mathbb{P}(\text{decoded payload} = P) = 0.$$

Proof. By Shannon’s channel coding converse (Shannon 1948), reliable communication (vanishing block error probability as $n \rightarrow \infty$) over a channel with capacity C requires $R < C$. For BSC($\frac{1}{2}$), $C = 0$. Hence no positive-rate code — including RS(n, k) (for which all n codeword bytes are erroneous since $e = n \gg t = \lfloor (n - k)/2 \rfloor$; for RS(255,127): $255 \gg 64$), any regular LDPC code (for which BP decoding achieves BER $\rightarrow \frac{1}{2}$ since $p = 0.5 \gg p^* \approx 0.084$ (Gallager 1962)), or Polar codes (for which all synthetic bit-channels polarize to pure noise since $\sum_i C^{(i)} = n \cdot C_{\text{BSC}}(\frac{1}{2}) = 0$ (Arikan 2009)) — can achieve reliable decoding on the sanitized channel. $\square$

Accuracy Stability Bound

Proposition 7 (Lipschitz Accuracy Bound): Let the task loss $\mathcal{L} : \mathcal{W} \rightarrow \mathbb{R}$ be L_{lip} -Lipschitz. Then:

$$|\Delta \text{Acc}| \leq L_{\text{lip}} \cdot \|W' - W\|_2 = L_{\text{lip}} \cdot \|W\|_2 \cdot D_{\text{rel}}(W, W').$$

Proof. Directly from Lipschitz continuity: $|\mathcal{L}(f_{W'}) - \mathcal{L}(f_W)| \leq L_{\text{lip}} \|W' - W\|_2 = L_{\text{lip}} \|W\|_2 \cdot D_{\text{rel}}(W, W')$. In the completed experiments, worst-case $D_{\text{rel}} = 5.4 \times 10^{-3}$ at $x = 19$. Spectral-norm analyses of BERT-scale models (Sakai and Imaizumi 2026) give $L_{\text{lip}} = O(10^2)$, yielding expected accuracy degradation consistent with the sub-1% changes reported in Tables 2 and 3 of the main paper. $\square$

Robustness under Imperfect Randomness

Lemma 8 (Capacity under Imperfect Randomness): Suppose the induced channel satisfies $\Pr(\hat{P}_{i,k} = P_{i,k}) = \frac{1}{2} + \epsilon$ for small $\epsilon > 0$. Then the effective capacity is bounded by:

$$C_{\text{eff}} \leq 1 - H_b(\frac{1}{2} + \epsilon) = O(\epsilon^2).$$

Proof. The binary entropy function achieves its maximum of 1 at $p = \frac{1}{2}$. Taylor expansion around $p = \frac{1}{2}$ gives:

$$H_b(\frac{1}{2} + \epsilon) = 1 - \frac{2}{\ln 2} \epsilon^2 + O(\epsilon^4).$$

Hence $C_{\text{eff}} \leq 1 - H_b(\frac{1}{2} + \epsilon) = O(\epsilon^2)$. This shows that even modest deviations from perfect uniformity still yield near-zero capacity, reinforcing the robustness of GRAYSHIELD beyond the idealized PRF model. $\square$

Threat Model

We consider a supply-chain adversary who embeds a binary payload into the mantissa LSBs of FP32 weights of a pretrained Transformer distributed through a public repository. The adversary has full white-box access to the model prior to distribution; the defender applies post-training sanitization before publication and has no access to the training data.

Target parameter set and hiding capacity We target the *encoder blocks* of the Transformer backbone: attention (Q/K/V/O) and feed-forward (MLP) weight matrices, which contain the dominant fraction of trainable parameters. Embedding layers and task-specific classifier heads are excluded by default. The hiding capacity at overwrite depth x is:

$$C(\theta_{\text{target}}, x) = \sum_{w \in \theta_{\text{target}}} \text{numel}(w) \cdot x, \quad x \in [1, 23]. \quad (6)$$

Multi-precision threat surface The implemented evaluation targets FP32. The analysis below extends the threat surface to deployment-grade precision formats; empirical validation of these formats is left as future work.

FP16 and BF16. FP16 (1-5-10) and BF16 (1-8-7) reduce mantissa depth to $m_{\text{FP16}} = 10$ and $m_{\text{BF16}} = 7$ bits, giving adjusted capacity $C_F = \sum_w \text{numel}(w) \cdot \min(x, m_F)$. For BERT-base encoder blocks at $x = 5$ in FP16, $C_{\text{FP16}} \approx 53$ MB, sufficient for both evaluated payloads.

INT8. Each INT8 weight embeds $x \leq 2$ bits at quantization error $\delta_i = q \cdot \text{scale}_\ell$ where $q \leq 1/128$, a negligible stealth cost at $x = 1$.

Three-tier attacker knowledge hierarchy

Definition 9 (Attacker Tiers):

- **Tier 1 (Black-box).** Knows architecture and parameter count; uses standard LSB substitution without knowledge of the defender’s method. Corresponds to NAIVE.
- **Tier 2 (Gray-box).** Knows a sanitization step of unspecified type will be applied; responds with redundancy coding (REPEAT3/REPEAT5) or burst-resilient interleaving.
- **Tier 3 (White-box).** Knows the sanitization algorithm (GRAYSHIELD) but not the seed s . Strongest realistic attacker.

Notation and Abbreviations

Table 4: Consolidated notation and abbreviations used throughout the paper.

Symbol / Abbrev.	Definition
<i>Weight-space and attack symbols</i>	
$W, \mathcal{W}$	Weight tensor; full model weight space
x	LSB overwrite depth ($1 \leq x \leq 23$ for FP32)
m_x	Bitmask ($1 \ll x$) $- 1$; selects lowest x bits
u_i	uint32 bitcast of weight w_i
N	Total number of target parameters
θ_{target}	Target parameter set (encoder blocks)
$C(\theta, x)$	Storage capacity Nx bits
C_{eff}	Effective Shannon capacity after sanitization
<i>Evaluation metrics</i>	
BitAcc	Bit accuracy $\frac{1}{B} \sum_i \mathbf{1}[\hat{p}_i = p_i]$ (lower better)
RR	Recovery Reduction $(\text{Rec}_{\text{pre}} - \text{Rec}_{\text{post}}) \times 100\%$
D_{rel}	Relative ℓ_2 distance $\ W' - W\ _2 / \ W\ _2$
W_1	1-Wasserstein distance between weight distributions
ΔAcc	Accuracy drop $\text{Acc}_{\text{clean}} - \text{Acc}_{\text{defended}}$
HD	Hamming distance (higher better for defense)
B	Payload length in bits
<i>GrayShield-specific symbols</i>	
s	Secret owner seed ($ s = 128$ bits)
r	Per-run salt (domain separation)
$H(\cdot; s)$	HMAC-SHA256 PRF keyed with s
$\mathcal{O}_\ell$	Per-tensor phase offset
$g_i^{(\ell)}$	Gray-code overwrite value at index i , tensor ℓ

continued on next page

Symbol / Abbrev.	Definition
<i>Information-theoretic symbols</i>	
$\text{BSC}(p)$	Binary symmetric channel with crossover probability p
$H_b(p)$	Binary entropy $-p \log_2 p - (1-p) \log_2 (1-p)$
$C_{\text{BSC}}(p)$	BSC capacity $1 - H_b(p)$ (zero at $p = 1/2$)
$\mathcal{I}(P; \hat{P})$	Mutual information between payload P and extracted bits $\hat{P}$
<i>Precision formats</i>	
FP32	32-bit single-precision (1 sign, 8 exponent, 23 mantissa)
FP16	16-bit half-precision (1-5-10)
BF16	Brain Float 16 (1-8-7)
INT8	8-bit signed integer with per-tensor scale
<i>Coding theory terms</i>	
ECC	Error-Correcting Code
RS	Reed–Solomon code (MDS over $\text{GF}(2^8)$)
LDPC	Low-Density Parity-Check code (BP decoding)
Polar	Polar code (SCL decoding)
<i>Defense abbreviations</i>	
GS	GrayShield (proposed)
PM	PatternMask: fixed-pattern LSB overwrite
RF	RandomFlip: per-bit Bernoulli flip
FT	FineTune: fine-tuning on clean samples
GN	GaussNoise: additive Gaussian noise
PTQ	Post-Training INT8 Quantization
SWP	Selective Weight Perturbation
<i>Models and datasets</i>	
BERT	BERT-base (110M params, SST-2)
RoBERTa	RoBERTa-base (125M params, SST-2)
ViT	ViT-Base/16 (86M params, CIFAR-10)
Swin	Swin-Tiny (29M params, CIFAR-10)
SST-2	Stanford Sentiment Treebank (binary)
CIFAR-10	10-class image benchmark

Formal Definitions

Definitions are ordered from the attack primitive upward to the defense guarantee.

Definition 10 (Payload): A *payload* is a binary string $\mathbf{p} = (p_1, \dots, p_B) \in \{0, 1\}^B$ of length $B \in \mathbb{N}$, drawn from an arbitrary distribution $\mathcal{P}$ over $\{0, 1\}^B$. The analysis in Section holds for every $\mathcal{P}$, including adversarially chosen fixed strings.

Definition 11 (Covert Channel): The *LSB covert channel* induced by the target parameter set θ_{target} , overwrite depth x , and model weights $W \in \mathcal{W}$ is the tuple $(\mathcal{E}, \mathcal{D}, \mathcal{W}, \{0, 1\}^C)$, where $C = C(\theta_{\text{target}}, x)$ is the storage capacity (Eq. 6), the *encoder* $\mathcal{E} : \mathcal{W} \times \{0, 1\}^C \rightarrow \mathcal{W}$ produces a weight tensor $W' = \mathcal{E}(W, \mathbf{p})$ by writing payload bits into the x mantissa LSBs of target weights, and the *decoder* $\mathcal{D} : \mathcal{W} \rightarrow \{0, 1\}^C$ recovers the payload by reading back those same bit positions: $\mathcal{D}(W') = \mathbf{p}$ with probability one in the noiseless pre-defense setting.

Definition 12 (Storage Capacity vs. Shannon Capacity): The *storage capacity* $C(\theta_{\text{target}}, x) = Nx$ (Eq. 6) is the number of payload bits the encoder can embed in N target weights at depth x ; it is a counting quantity. The *effective Shannon capacity* $C_{\text{eff}}(T) = \mathcal{I}(P; \hat{P})$ (Eq. 5) is the maximum mutual information between the embedded payload P and the extracted sequence $\hat{P}$ after sanitization map T ; it is an information-theoretic quantity measured in bits per channel use. The goal of GRAYSHIELD is to reduce C_{eff} to zero while leaving $C(\theta_{\text{target}}, x)$ unchanged.

Definition 13 (Post-Training Sanitization Map): A *post-training sanitization map* is a function $T : \mathcal{W} \times \mathcal{K} \rightarrow \mathcal{W}$, where $\mathcal{W} = \prod_{\ell} \mathbb{R}^{N_{\ell}}$ is the model weight space and $\mathcal{K}$ is a key space, satisfying:

1. **Data-free:** $T(W, k)$ does not require access to any training or validation data.
2. **Weight-only:** T modifies only the weight tensor W ; model architecture, tokenizer, and configuration files are unchanged.
3. **$O(N)$ complexity:** The runtime of T is linear in the number of target parameters $N = \sum_{\ell} N_{\ell}$.

GRAYSHIELD instantiates T with $\mathcal{K} = \{0, 1\}^{|\mathcal{s}|}$ (the seed space), deterministically as described in Section .

Definition 14 (Pseudorandom Function): A *pseudorandom function* (PRF) is a keyed function $H : \{0, 1\}^* \times \{0, 1\}^{|s|} \rightarrow \{0, 1\}^\lambda$ such that no probabilistic polynomial-time adversary can distinguish $H(\cdot, s)$ from a uniformly random function $f : \{0, 1\}^* \rightarrow \{0, 1\}^\lambda$ with non-negligible advantage, for a uniformly sampled key $s \leftarrow \{0, 1\}^{|s|}$. Formally, for all PPT distinguishers $\mathcal{A}$:

$$\left| \mathbb{P}_s[\mathcal{A}^{H(\cdot, s)} = 1] - \mathbb{P}_f[\mathcal{A}^{f(\cdot)} = 1] \right| \leq \text{negl}(|s|). \quad (7)$$

GRAYSHIELD instantiates H as HMAC-SHA256, which satisfies Eq. (7) under standard assumptions on the SHA-256 compression function (Bellare, Canetti, and Krawczyk 1996).

Definition 15 (Effective Channel Capacity): Let $T : \mathcal{W} \times \mathcal{K} \rightarrow \mathcal{W}$ be a sanitization map and let $\hat{P} = \mathcal{D}(T(W, k))$ denote the payload extracted from the sanitized weights. The *effective channel capacity* of the covert channel under defense T with key k is

$$C_{\text{eff}}(T, k) = \sup_{\mathcal{P}} \mathcal{I}_{\mathcal{P}}(P; \hat{P}), \quad (8)$$

where the supremum is taken over all payload distributions $\mathcal{P}$ on $\{0, 1\}^B$, and $\mathcal{I}_{\mathcal{P}}$ denotes mutual information under that distribution. Proposition 4 establishes that $C_{\text{eff}}(\text{GRAYSHIELD}, s) = 0$ for all $s \in \mathcal{K}$ unknown to the attacker under the idealized PRF model.

Definition 16 (Payload Recovery): Given an embedded payload $\mathbf{p} \in \{0, 1\}^B$ and a decoded sequence $\hat{\mathbf{p}} \in \{0, 1\}^B$ obtained after applying defense T , the *payload recovery rate* is

$$\text{Rec}(\mathbf{p}, \hat{\mathbf{p}}) = \frac{1}{B} \sum_{i=1}^B \mathbf{1}[\hat{p}_i = p_i] = \text{BitAcc}(\mathbf{p}, \hat{\mathbf{p}}). \quad (9)$$

The *Recovery Reduction* is then $\text{RR} = (\text{Rec}_{\text{pre}} - \text{Rec}_{\text{post}}) \times 100\%$, where $\text{Rec}_{\text{pre}} = \text{BitAcc}(\mathbf{p}, \mathcal{D}(\mathcal{E}(W, \mathbf{p})))$ is the pre-defense recovery rate (effectively 1.0 for aligned injection at $x \leq 16$) and $\text{Rec}_{\text{post}} = \text{BitAcc}(\mathbf{p}, \mathcal{D}(T(\mathcal{E}(W, \mathbf{p}), k)))$ is the post-defense recovery rate.

Definition 17 (Stealth): An embedding $\mathcal{E}(W, \mathbf{p})$ is $(\epsilon_\ell, \epsilon_c, \epsilon_a)$ -*stealthy* if the following three conditions hold simultaneously:

$$D_{\text{rel}}(W, \mathcal{E}(W, \mathbf{p})) \leq \epsilon_\ell \quad (\text{perturbation magnitude}), \quad (10)$$

$$1 - \text{Sim}_{\text{cos}}(W, \mathcal{E}(W, \mathbf{p})) \leq \epsilon_c \quad (\text{directional preservation}), \quad (11)$$

$$|\text{Acc}(f_W) - \text{Acc}(f_{\mathcal{E}(W, \mathbf{p})})| \leq \epsilon_a \quad (\text{utility preservation}). \quad (12)$$

In the present study, the empirical capacity boundary of $(10^{-4}, 2.6 \times 10^{-7}, 0.01)$ -stealth is confirmed at $x=16$ across all four model presets, degrading to $(8.5 \times 10^{-2}, 3.5 \times 10^{-3}, 0.01)$ -stealth at $x=23$.

Definition 18 (Perturbation Field): Let $W^{(\ell)} \in \mathbb{R}^{N_\ell}$ be a target parameter tensor and let $W^{(\ell)'} = T_\ell(W^{(\ell)})$ denote its sanitized version. The *perturbation field* of sanitization method M on tensor ℓ is the function $\Delta^{(M, \ell)} : \{0, \dots, N_\ell - 1\} \rightarrow \mathbb{R}_{\geq 0}$ defined by

$$\Delta_i^{(M, \ell)} = |w_i^{(\ell)'} - w_i^{(\ell)}| = g_i^{(\ell)} \cdot 2^{-23} \cdot 2^{e_i - 127}, \quad (13)$$

where $g_i^{(\ell)}$ is the overwrite value at position i and the indices follow the canonical row-major flattening order of the parameter tensor. The total perturbation field $\Delta^{(M)}$ of the full model is the concatenation of $\Delta^{(M, \ell)}$ across all target tensors in flattening order.

Definition 19 (Total-Variation Smoothness): The perturbation field $\Delta^{(M)}$ of method M is τ -*total-variation smooth* if

$$E_{\text{TV}}(M) = \sum_{i=0}^{N-2} |\Delta_{i+1}^{(M)} - \Delta_i^{(M)}| \leq \tau. \quad (14)$$

A method with lower E_{TV} introduces more slowly varying perturbations in the flattened weight order, reducing (i) the high-frequency spectral energy detectable by frequency-domain anomaly scanners, and (ii) the risk of exciting sharp curvature directions in the task-loss Hessian. The Gray-code property $\text{Ham}(g_i^{(\ell)}, g_{i+1}^{(\ell)}) \leq 1$ implies $\tau_{\text{GRAYSHIELD}} \leq N \cdot 2^{-23} \cdot \mathbb{E}[2^{e_i}]$, whereas PM with a period-2 pattern achieves $\tau_{\text{PM}} = \Theta(N \Delta_{\text{max}})$, giving $\tau_{\text{GS}} \ll \tau_{\text{PM}}$.

Definition 20 (Memoryless Covert Channel): The covert channel $(\mathcal{E}, \mathcal{D}, \mathcal{W}, \{0, 1\}^C)$ is *memoryless with respect to sanitization* T if, for all $i \neq j$, the random variables $\hat{P}_i = \mathcal{D}(T(W, k))_i$ and $\hat{P}_j = \mathcal{D}(T(W, k))_j$ are conditionally independent given $P = (P_1, \dots, P_B)$:

$$\hat{P}_i \perp \hat{P}_j \mid P, k \quad \forall i \neq j. \quad (15)$$

This condition holds for GRAYSHIELD because $g_i^{(\ell)} = ((i + \mathcal{O}_\ell) \oplus ((i + \mathcal{O}_\ell) \ggg 1)) \& m_x$ is a deterministic function of i alone (given $\mathcal{O}_\ell$), with no dependence on $g_j^{(\ell)}$ for $j \neq i$. The post-defense extracted bits are therefore independent across positions, justifying the memoryless BSC model.

Definition 21 (Chance-Level Recovery): For a binary payload channel with crossover probability p , *chance-level recovery* is the expected bit accuracy under the maximum-entropy noise distribution:

$$\text{Rec}_{\text{chance}} = \mathbb{P}(\hat{P}_i = P_i) = 1 - p \xrightarrow{p \rightarrow 1/2} \frac{1}{2}. \quad (16)$$

Proposition 4 establishes that GRAYSHIELD drives $p \rightarrow 1/2$ in the idealized PRF model, so $\text{Rec}_{\text{post}} \rightarrow 1/2$ and $\text{RR} \rightarrow 50\%$. “Near-chance sanitization” in our empirical results refers to $\text{Rec}_{\text{post}} \in [0.49, 0.51]$, corresponding to a channel crossover probability indistinguishable from $1/2$ at the measured sample sizes.

Extended Trade-off Analysis

Figure 6 provides an extended view of the security–utility trade-off across defenses under both baseline and adaptive attacker settings. We place this figure in the appendix because the full configuration-level visualization is informative but visually dense relative to the main paper figures.

The figure complements the main RQ3 analysis by showing the distribution of recovery reduction (RR) and accuracy degradation across all evaluated model, payload, and attacker configurations. It confirms that the high-level ordering observed in the main results remains stable across attacker variants, while also exposing important differences in variance, operating-point consistency, and deployment robustness among defenses.

Multi-Precision Threat Extension

The main evaluation targets FP32. FP16 (1-5-10) and BF16 (1-8-7) reduce mantissa depth to $m_{\text{FP16}} = 10$ and $m_{\text{BF16}} = 7$ bits respectively, giving adjusted hiding capacity $C_F = \sum_w \text{numel}(w) \cdot \min(x, m_F)$. For BERT-base encoder blocks at $x = 5$ in FP16, $C_{\text{FP16}} \approx 53$ MB, sufficient for both evaluated payloads. Proposition 4 extends to these formats unchanged, since the Gray-code overwrite applies identically by clamping $x \leq m_F$. Each INT8 weight supports $x \leq 2$ bits at quantization error $\delta_i = q \cdot \text{scale}_\ell$ with $q \leq 1/128$, a negligible stealth cost at $x = 1$. Empirical validation of reduced-precision formats is left as future work.

Bit Manipulation in FP32 Weights

A float32 value stores sign (1 bit), exponent (8 bits), and mantissa (23 bits). All attack and defense operations act exclusively on mantissa LSBs, preserving sign and exponent unchanged.

Let $u = \text{bitcast}(w) \in \{0, 1\}^{32}$ be the uint32 view of weight w and $m = (1 \ll x) - 1$ the LSB mask. The three core operations are:

- *Extraction*: $u \& m$.
- *Overwrite / sanitization*: $u' = (u \& \sim m) \mid b$, $b \in [0, 2^x - 1]$.
- *Random flip*: $u' = u \oplus r$ within the masked region, where r is sampled from $\text{Bernoulli}(p)$ independently per bit.

All three are fully vectorizable via `Tensor.view(torch.int32)` on contiguous FP32 parameters, completing in $O(N)$ time with $O(1)$ auxiliary state per tensor.

Budget Definition

The *bit-level budget* fixes the total number of modified bits B , implying $K = \lfloor B/x \rfloor$ modified floats. The *float-level budget* modifies K float elements, touching $K \cdot x$ bits. We use a bit-level budget throughout the paper for fairness across different values of x .

Additional Fidelity Metrics

GRAYSHIELD does not yield the single largest average Hamming distance; that distinction is held by PM by a negligible margin. However, GRAYSHIELD achieves essentially the same level of payload corruption while preserving substantially better weight fidelity: its average W_1 is about $25\times$ smaller than PM and about $42\times$ smaller than PTQ. The main empirical advantage of GRAYSHIELD is therefore near-maximal corruption with markedly cleaner distribution-level behaviour.

Reproducibility

The anonymized replication package (Anonymous Authors 2026) contains the code, configuration grids, figure-generation scripts, curated payload data, dataset metadata, structured result artifacts, and provenance information needed to reproduce the main figures and tables in this paper.

The curated payload data include the two representative payloads used in the completed main-paper runs—the low-entropy JavaScript sample `c37...a88.js.malware` and the high-entropy executable `5704...6d0.exe.malware`—together with four supplementary payloads and a metadata manifest recording file type, size, entropy, and paper role. The payload bundle is included in the anonymized replication package.

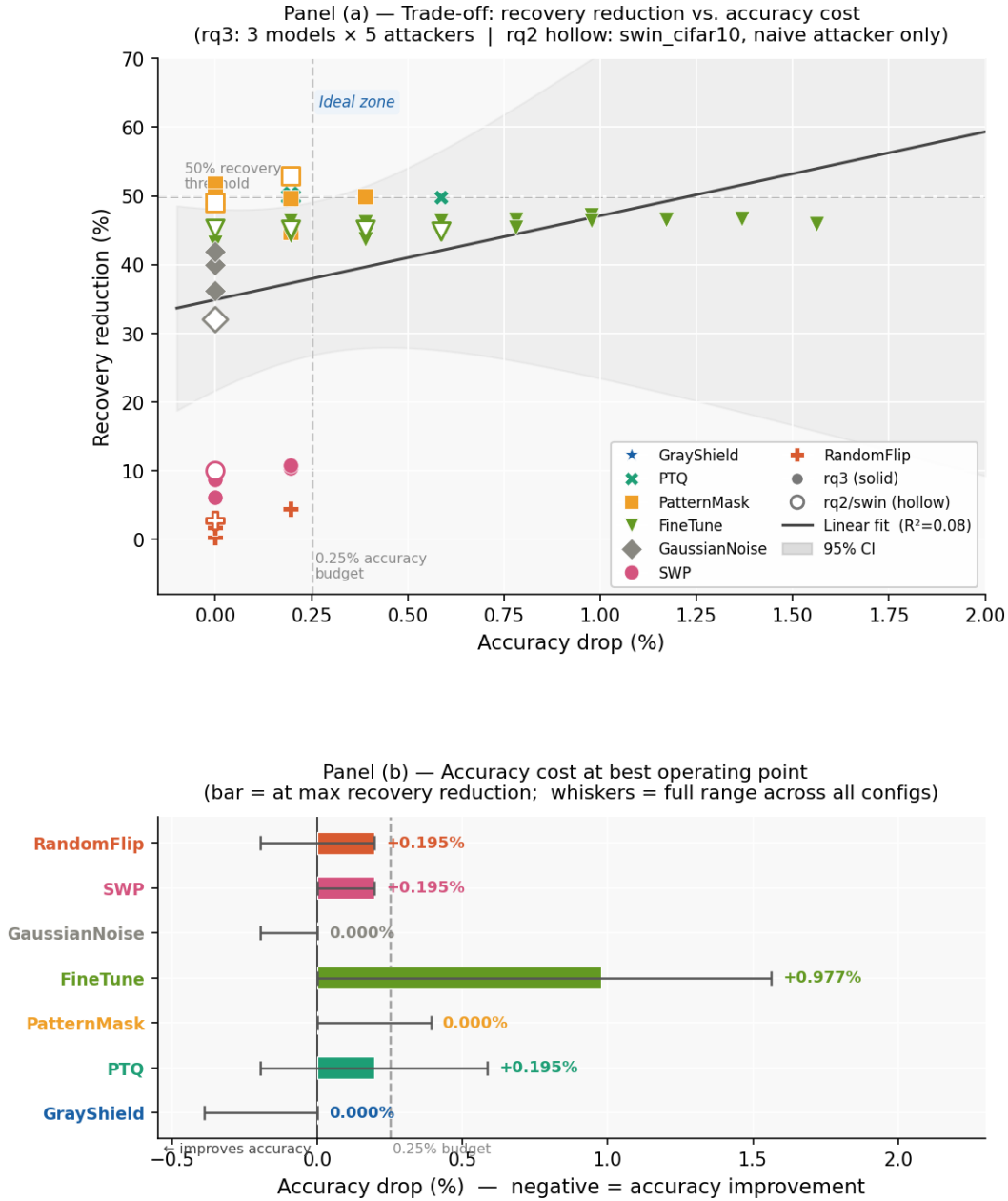


Figure 6: **Security–utility trade-off across defenses.** **Panel (a)** plots recovery reduction (RR) versus accuracy drop for all defense configurations. Solid markers denote configurations evaluated under the RQ3 adaptive attacker setting; hollow markers indicate the additional `swin_cifar10` configuration from RQ2. Hollow markers show the additional `swin_cifar10` configuration from RQ2. The horizontal dashed line marks the desired 50% chance-level recovery threshold, while the vertical dashed line indicates the practical 0.25% accuracy budget. GS consistently occupies the upper-left “ideal zone,” achieving near-chance RR with minimal accuracy degradation and low variance. FT achieves competitive RR only at substantially higher utility cost, whereas GN, SWP, and RF fail to provide effective sanitization. The weak global trend ($R^2 = 0.08$) indicates that stronger sanitization does not necessarily require larger accuracy degradation. **Panel (b)** shows the accuracy drop at each defense’s best operating point, with whiskers indicating the full observed range across configurations. GS achieves near-optimal RR at effectively zero accuracy cost while maintaining substantially tighter variability than competing high-RR defenses.

The repository also includes a structured results bundle under `release_results/`. This bundle separates the refreshed RQ1 rerun from the completed RQ2–RQ4 results, and includes sanitization configuration files, attacker encoder/decoder

implementations, figure-generation utilities, and a merged provenance manifest tracing each published result back to its source run.

Seeds: run seed 1; evaluation seed 42; $n_{\text{eval}} = 512$ per run.

Software: Python 3.13, PyTorch 2.9.1, Transformers 5.0.0.

Hardware: Ubuntu 22.04.5; $2 \times$ Intel Xeon Gold 6430 (128 logical cores); 503 GiB RAM; $2 \times$ NVIDIA RTX PRO 6000 Blackwell (96 GB VRAM each).