

Structural Collapse of Safety-Critical Weight Protection Scorers Under Group-Wise Post-Training Quantization

Jianchang Su

University of Connecticut
jianchang.su@uconn.edu

Wei Zhang*

University of Connecticut
wei.13.zhang@uconn.edu

Abstract

Post-training quantization (PTQ) of instruction-tuned LLMs to 3 bits often destroys safety alignment, motivating *safety-protection* methods that keep a small budget of “safety-critical” weight blocks in full precision. We give a theoretical account of the many scoring rules proposed to choose those blocks. We prove that under symmetric group-wise PTQ the Taylor (SSMP) scorer concentrates to the gradient magnitude times a per-group constant; the gradient-times-weight (SNIP) scorer is empirically near-identical, so the entire compound family $\alpha|g \cdot w| + \beta|g \cdot \delta w|$ collapses to a single block ranking (worst-case Spearman 0.9971), and only pure magnitude or pure Fisher selects different blocks. We further show that recovered refusal behavior under small budgets is *decoupled* from the underlying safety mechanism: refusal can be restored while capability, adversarial robustness, and over-refusal balance stay at the unprotected level; at 3 bits heuristic attack metrics conflate incoherence with harm; and on weakly-aligned models protection raises classifier-judged attack success above the unprotected floor. Experiments on five open-weight 8–9B models across 110 conditions, including classifier-judged GCG and PAIR attacks and activation-level analysis, confirm every prediction; a four-category classification predictable from one baseline metric explains the cross-model variation. Code and configurations: <https://github.com/jc-su/Safepress>.

1 Introduction

Quantizing an instruction-tuned LLM to 3 bits under post-training quantization (PTQ) can drop its refusal rate against harmful prompts dramatically. On Llama-3.1-8B-Instruct, refusal against the HarmBench set falls from 67.7% at FP16 to 36.7% under uniform 3-bit symmetric group-wise quantization.

* Corresponding author.

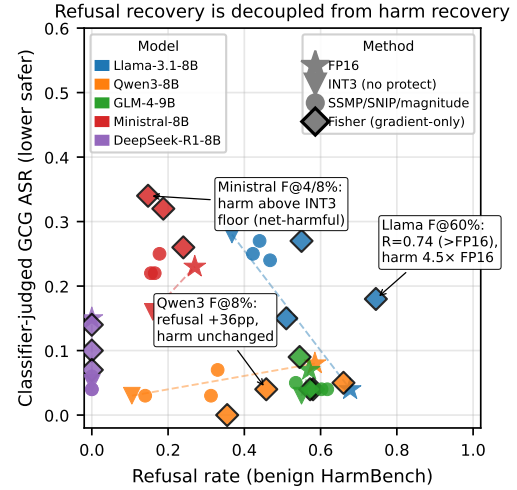


Figure 1: **The central finding (Proposition 3).** Classifier-judged GCG attack ASR (y , lower safer) vs. HarmBench refusal rate (x , higher safer) for the 40 main-panel conditions at 3-bit. Dashed connectors mark each model’s recovery path from unprotected INT3 to FP16: a method that restored classifier-judged robustness would move protected points along them. Protected points move horizontally (Qwen3 Fisher@8%: refusal +36pp, harm unchanged), stall far from FP16 (Llama Fisher@60%: refusal above FP16, harm 4.5× FP16), or move up (Ministral Fisher@4/8%: net-harmful). Refusal recovery is decoupled from recovery of adversarial robustness.

The same model’s attack success rate (ASR) under transferred Greedy Coordinate Gradient (GCG) suffixes (Zou et al., 2023) rises from 6% to 53% on the heuristic non-refusal metric, and from 4% to 28% under classifier-judged harm; its WikiText-2 PPL triples from 8.7 to 27.1. The phenomenon is documented across multiple instruction-tuned model families and motivates a growing literature on *safety-aware quantization*: mixed-precision approaches that identify a small set of “safety-critical” weights to keep in full precision while quantizing the rest (Chen et al., 2025; Al Hakim et al., 2026), objective-level safety-preserving PTQ (Wee

et al., 2025), and quantized-gradient safety analysis (Padakandla et al., 2025).

These approaches differ in their scoring rule. SNIP-style methods (Single-shot Network Pruning; Lee et al., 2019) score a block by $\sum_j |g_j w_j|$ where g is the gradient of a safety loss with respect to the weights. Taylor-style methods (Safety-Selective Mixed Precision, SSMP) score by $\sum_j |g_j \delta w_j|$, the element-wise first-order drift contribution. Fisher-based methods (CWP) use g^2 (or a difference of safety and general-purpose Fishers). Magnitude-based methods use $|w|$ alone.

Despite the diversity of proposed scorers, three observations recur empirically: **(1)** compound scoring rules involving the weight-magnitude factor produce nearly indistinguishable block rankings under group-wise PTQ; **(2)** “recovered” refusal under small-budget protection often fails to transfer to adversarial probes (GCG (Zou et al., 2023), PAIR (Chao et al., 2025)) or capability metrics (MMLU (Hendrycks et al., 2021)); **(3)** on some models, gradient-magnitude protection *worsens* adversarial robustness relative to unprotected uniform quantization.

We argue that these observations are structural consequences of the interaction between symmetric group-wise PTQ and gradient-based scoring rules: the refusal behavior that small-budget protection recovers is decoupled from the safety it is meant to preserve, and a different compound scorer leaves this unchanged, because the entire compound family selects the same blocks. Two theorems and one proposition formalize this argument, validated on a 5-model, 110-cell panel.

Contributions.

- **Proposition 3 (Decoupling, the central finding):** At bit-widths ≤ 3 with protection budget $\leq 8\%$, refusal-rate restoration is decouplable from capability, adversarial robustness, and over-refusal balance: classifier-judged evaluation on two attack families (GCG, PAIR) and an activation-level analysis show that protection restores refusal behavior while classifier-judged robustness and the refusal representation stay at their unprotected levels, and on weakly-aligned models protection is net-harmful.
- **Theorem 2 (Structural Collapse):** Under group-wise PTQ the SSMP scorer $\sum_j |g_j \delta w_j|$ concentrates to the gradient

ℓ_1 norm times a per-group scale; SNIP $\sum_j |g_j w_j|$ is empirically near-identical, so the entire compound family $\alpha|g \cdot w| + \beta|g \cdot \delta w|$ ranks blocks as one equivalence class (worst-case Spearman 0.9971; protect-set Jaccard 0.89–0.97), invariant to block size, quantizer structure, and scale. Only dropping the gradient or the magnitude factor (pure $|w|$ or pure $\|g\|^2$) escapes.

- **Theorem 1 (Drift Bound):** A first-order Cauchy–Schwarz upper bound on safety-loss change under PTQ that orders drift severity across bit-widths (pooled Spearman 0.834; 2-bit catastrophe separated by 40–70 $\times$).
- An empirical four-category classification of models, parametrized by FP16 GCG ASR and bit-width, predicting which models can benefit from protection at all.

The contribution is analysis rather than a new scoring rule: the results explain why the existing family clusters, and show that claimed recovery requires classifier-based evaluation in addition to refusal rate.

2 Related Work

Post-training quantization. AWQ (Lin et al., 2024), GPTQ (Frantar et al., 2023), and bitsandbytes NF4 (Dettmers et al., 2023) represent the dominant deployment-grade PTQ approaches at ≤ 4 bits; salience-driven mixed-precision schemes (SliM-LLM (Huang et al., 2025), LLM-MQ (Li et al., 2023)) allocate precision by output-salience rather than safety. Target-aware calibration instead selects calibration examples to preserve deployment-specific uncertainty behavior (Yang et al., 2026); this is complementary to our focus on selecting weight blocks to preserve safety behavior. We use simulated symmetric group-wise quant-dequant (the arithmetic underlying NF4 prior to storage compression) for analytical tractability; Theorem 2 applies to any quantizer expressible as group-wise scale $\times$ bounded perturbation, verified empirically for four quantizer structures (App. E).

Safety-aware quantization. Q-resafe (Chen et al., 2025) scores by SNIP $\sum_j |g_j w_j|$ and adds a LoRA-DPO (Hu et al., 2022; Rafailov et al., 2023) post-quantization fine-tune. CWP (Al Hakim et al., 2026) scores by $\mathbf{I}_{\text{safe}} - \beta \mathbf{I}_{\text{gen}}$ (a difference of safety and general Fishers) at $\sim 60\%$ protection

budget. Contrastive Alignment Quantization (Wee et al., 2025) changes the PTQ objective to preserve safety alignment. SafeQuant (Padakandla et al., 2025) uses quantized-gradient inspection to analyze unsafe-behavior patterns. Our Theorem 2 implies that the *ranking* produced by Q-resafe-SNIP, SSMP-Taylor, and non-negative combinations of those two gradient-weighted scorers is equivalent up to within-block magnitude dispersion; pure magnitude and pure Fisher escape by dropping one of the two factors.

Prior work in this line proposes scorers; none analyzes their equivalence. The closest methodological cousins are saliency-equivalence results for pruning at initialization: under infinite-width regularity conditions, GraSP becomes asymptotically rank-equivalent to SNIP (Pham et al., 2026); GraSP introduced gradient-flow preservation (Wang et al., 2020), while SynFlow derived saliency conservation laws (Tanaka et al., 2020). These results concern weight *removal* and follow from the loss gradient at initialization. Ours is quantization-specific: the group-wise error structure ($\delta w \propto \text{per-group scale} \times \text{bounded residual}$) makes the quantizer-aware $|g \cdot \delta w|$ scorer collapse onto the $|g \cdot w|$ family. The theory also explains a published empirical fact: CWP reports that its Fisher-family score beats SNIP, and Fisher is exactly the scorer that escapes the collapse.

Adjacent model-integrity systems. Safety degradation under quantization is one instance of the broader problem of preserving model integrity across the learning and deployment lifecycle. At training time, BadWindtunnel identifies poisoned examples through confidence variance in a high-noise simulated environment (Zhang et al., 2025a), while BadLLM-TG uses an LLM-based trigger generator to support trigger inversion and adversarial retraining (Zhang et al., 2026a). Related anomaly-detection work explicitly separates difficult clean samples from anomalous contamination (Zhang et al., 2026b). At the systems level, FastMatch targets distributed-training input-pipeline efficiency (Su et al., 2024), and TrustWeave measures and attests the integrity of LLM services spanning multiple clouds (Su et al., 2026). These works address distinct threat models and system layers; we cite them as complementary integrity mechanisms rather than as safety-aware PTQ baselines.

Refusal mechanisms. Arditi et al. (2024) show refusal in instruction-tuned models is mediated by a low-rank direction in the residual stream. Subsequent work documents richer structure: multi-dimensional refusal concept cones with representational independence (Wollschläger et al., 2025), cross-lingual universality of the refusal direction (Wang et al., 2025), nonlinear refusal geometry (Hildebrandt et al., 2025), and refusal induced by knowledge-scope limitation (Cao, 2024). Our activation-level analysis (App. M) finds the single-direction separability essentially unchanged under 3-bit PTQ while refusal behavior collapses, consistent with the multi-dimensional and nonlinear view: a preserved linear direction does not guarantee preserved behavior. Proposition 3 explains why refusal behavior can survive quantization while adversarial robustness degrades.

3 Setup and Notation

Problem setting. A deployer wishes to quantize an instruction-tuned LLM to ≤ 3 bits to fit memory constraints, while preserving its safety alignment (its tendency to refuse harmful requests). Mixed-precision *safety protection* keeps a small budget β of “safety-critical” weight blocks in full precision and quantizes the rest. Our analysis asks two questions: (Q1) do the proposed scoring rules for identifying safety-critical blocks actually differ, and (Q2) when a protection scheme “recovers” refusal behavior, has it recovered the underlying safety mechanism?

Model and blocks. We treat an LLM as a stack of Linear modules $\{W_\ell\}_{\ell=1}^L$, $W_\ell \in \mathbb{R}^{d_{\text{out}} \times d_{\text{in}}}$. Each W_ℓ is partitioned into *blocks* of $K = 64$ consecutive output rows; we write $w_b \in \mathbb{R}^{K \cdot d_{\text{in}}}$ for the flattened weights of block b and g_b for the corresponding safety-loss gradient. All scoring and protection operate at block granularity.

Definition 1 (Symmetric group-wise quantizer). For a target bit-width and group size $G=128$, partition each weight row into groups. For group $\mathcal{G}$, let the scale be $s_{\mathcal{G}} = \max_{j \in \mathcal{G}} |w_j| / q_{\text{max}}$ with $q_{\text{max}} = 2^{\text{bit-width}-1} - 1$. The quantize–dequantize map is

$$Q(w_j) = s_{\mathcal{G}} \cdot \text{clip}(\text{round}(w_j / s_{\mathcal{G}}), -q_{\text{max}}, q_{\text{max}}), \quad (1)$$

with perturbation $\delta w_j = Q(w_j) - w_j$. Writing $r_j = w_j / s_{\mathcal{G}} - \text{round}(w_j / s_{\mathcal{G}}) \in [-\frac{1}{2}, \frac{1}{2}]$ for the *rounding residual*, we have $\delta w_j = -s_{\mathcal{G}} r_j$.

Definition 2 (Safety loss and scoring rules). Given harmful-intent prompts $\mathcal{D}_{\text{safe}}$ and a refusal template r^* , the safety loss is $L_{\text{safe}}(W) = -\mathbb{E}_{x \sim \mathcal{D}_{\text{safe}}} \log p_W(r^* | x)$, with block gradient $g_b = \partial L_{\text{safe}} / \partial w_b$. We study four block scoring rules:

$$\text{SNIP: } s_b^{\text{snip}} = \sum_{j \in b} |g_j w_j| \quad (2)$$

$$\text{SSMP/Taylor: } s_b^{\text{ssmp}} = \sum_{j \in b} |g_j \delta w_j| \quad (3)$$

$$\text{Magnitude: } s_b^{\text{mag}} = \|w_b\|_1 \quad (4)$$

$$\text{Fisher: } s_b^{\text{fish}} = \|g_b\|_2^2 \quad (5)$$

Given budget β , the *protect set* P is the top- β fraction of blocks by score; weights in P stay FP16, the rest are quantized via Def. 1.

Assumption 1 (High-resolution rounding). For weights whose scaled fractional parts $\{w_j/s_g\}$ are equidistributed modulo 1 (the generic, non-lattice-aligned case at $b \leq 4$ bits), the rounding residuals r_j behave as i.i.d. uniform on $[-\frac{1}{2}, \frac{1}{2}]$, independent of the gradient g_j and of the weight value w_j . This is the standard quantization-noise model (Widrow et al., 1996).

Assumption 1 is the source of the structural-collapse phenomenon: it says the quantization *error direction* carries no information about which individual weights are large, only the per-group *scale* does. App. B validates it directly on the selected blocks (effective dimension $\geq 1.4 \times 10^3$ on every model, worst-case Lemma-1 fluctuation 1.6%), and the collapse survives correlated error-feedback residuals, where the i.i.d. premise fails (App. E).

4 Theoretical Analysis

4.1 Theorem 1: Drift Bound

Theorem 1 (Drift Bound). Let L_{safe} be twice differentiable in W and let δW be a PTQ perturbation with bounded operator norm. Then for the block decomposition above,

$$\Delta L_{\text{safe}} = \sum_b \langle g_b, \delta w_b \rangle + O(\|\delta W\|_2^2), \quad (6)$$

and

$$|\Delta L_{\text{safe}}| \leq \sum_b \|g_b\|_2 \cdot \|\delta w_b\|_2 + O(\|\delta W\|_2^2). \quad (7)$$

Proof sketch. First-order Taylor expansion of $L_{\text{safe}}(W + \delta W)$ around W gives the linear term

plus an $O(\|\delta W\|_2^2)$ remainder. Per-block Cauchy–Schwarz bounds each $|\langle g_b, \delta w_b \rangle|$, and the triangle inequality assembles the first-order block-sum; the remainder is the stated $O(\|\delta W\|_2^2)$ term, which the empirical fit (§6) confirms is small relative to the linear bound. $\square$

Implications. (i) The per-block drift magnitude is bounded by the SSMP scorer via the triangle inequality, $|\langle g_b, \delta w_b \rangle| \leq \sum_{j \in b} |g_j \delta w_j| = s_b^{\text{ssmp}}$, so greedily protecting the highest-SSMP blocks bounds the total drift; this is the justification for the Taylor/SSMP family. (ii) We validate the bound empirically by comparing the predicted block-sum $\sum_b |\langle g_b, \delta w_b \rangle|$ against measured $|\Delta L_{\text{safe}}|$ across bit-widths: the bound is a strict upper envelope that orders drift severity (pooled rank correlation 0.834 over 24 points; 40–70 $\times$ separation of the 2-bit catastrophe, §6).

4.2 Theorem 2: Structural Collapse

The SSMP scorer is the element-wise absolute sum $s_b^{\text{ssmp}} = \sum_{j \in b} |g_j \delta w_j|$ (Def. 2, matching the standard implementation). The central result is that under group-wise quantization this scorer concentrates to the gradient ℓ_1 norm scaled by the per-group constant; it loses the fine-grained dependence on individual weights. (Proof in App. A.)

Lemma 1 (SSMP score concentration). Under Def. 1–2 and Assumption 1, writing s_b for the (dominant) group scale of block b and $|\delta w_j| = s_b |r_j|$ with $|r_j|$ i.i.d., $\mathbb{E}|r_j| = \frac{1}{4}$,

$$s_b^{\text{ssmp}} = s_b \sum_{j \in b} |g_j| |r_j| = \frac{s_b}{4} \|g_b\|_1 (1 + o(1)), \quad (8)$$

where the concentration holds because s_b^{ssmp} is a sum of $n=Kd_{\text{in}}$ non-negative i.i.d.-scaled terms (law of large numbers, relative fluctuation $O(1/\sqrt{n})$). (The signed inner product $g_b \cdot \delta w_b$ is zero-mean and only converges in distribution; the absolute-sum scorer that is actually used concentrates.)

Theorem 2 (Structural Collapse). Under the assumptions of Lemma 1, the SSMP score concentrates to $\frac{s_b}{4} \|g_b\|_1$, the gradient ℓ_1 norm times the per-group scale $s_b = \max_{\mathcal{G}} |w|/q_{\text{max}}$. Each scorer in Def. 2 is characterized by which of two factors it retains, a gradient factor and a weight-magnitude factor; the consequences follow:

1. (**Collapse of two-factor scorers.**) SSMP ranks by $s_b \|g_b\|_1$; SNIP = $\sum_j |g_j| |w_j|$ ranks by

the gradient magnitudes weighted by per-coordinate $|w_j|$. Both are the gradient ℓ_1 mass weighted by a weight-magnitude statistic (the group scale s_b for SSMP, per-coordinate $|w_j|$ for SNIP). They coincide up to within-block weight-magnitude dispersion; when this dispersion is small the two rankings are identical. We do not assume this analytically; we verify it empirically, finding SSMP and SNIP near-perfectly rank-correlated (Spearman 0.997–1.000, §7).

2. (**Magnitude escape.**) Pure magnitude $\|w_b\|_1$ drops the gradient factor; it diverges from the two-factor class whenever the safety gradient is not comonotone with weight magnitude.
3. (**Fisher escape.**) Pure Fisher $\|g_b\|_2^2$ drops the weight-magnitude factor; it diverges whenever $\|g_b\|$ and weight magnitude are not comonotone across the top- β blocks.

Corollary 1 (Two-factor equivalence class). *The two gradient-weighted scorers $\{\sum |g w|, \sum |g \delta w|\}$ (SNIP, SSMP) rank blocks equivalently up to within-block weight-magnitude dispersion. The single-factor scorers, pure magnitude $\|w\|_1$ (gradient dropped) and pure Fisher $\|g\|_2^2$ (magnitude dropped), each diverge, in different directions.*

Implications. Q-resafe (SNIP+DPO) and SSMP (Taylor) select the *same* safety-critical blocks under group-wise PTQ up to negligible within-block dispersion; they differ only in their post-scoring step (retraining, budget). The corollary makes a sharp, directly testable block-selection prediction: SSMP and SNIP protect-sets should nearly coincide, while pure-magnitude and Fisher diverge, as we verify in §7. Eq. 8 also explains Theorem 1’s magnitude-tightness variation: the bound scale is $\propto s_b$, a per-group constant differing across models.

The equivalence class is moreover a *continuum*: every non-negative mixture $\alpha|g \cdot w| + \beta|g \cdot \delta w|$, including Q-resafe’s SNIP scoring, ranks blocks with worst-case Spearman 0.9971 to SNIP over an (α, β) grid on all five models (App. C). The equivalence is also independent of the analysis constants: the correlation is ≥ 0.998 at every block size $K \in \{16, \dots, 128\}$, tightening with K per the $O(1/\sqrt{K d_{\text{in}}})$ rate; G enters only through the per-group scale, which preserves the block ranking; and the collapse holds under NF4, asymmetric,

per-channel, and error-feedback quantizers and at 14B (Apps. D–F).

4.3 Proposition 3: Refusal-Mechanism Decoupling

Our third result is an empirical proposition (not an analytic theorem): we state it as a falsifiable claim and validate it by construction in §8.

Proposition 3 (Refusal-Mechanism Decoupling). *At bit-widths ≤ 3 and protection budget $\beta \leq 8\%$, there exist protection schemes whose protected-and-quantized model has refusal-rate against benign harmful-intent probes substantially restored toward FP16, while at least one of the following remains at the unprotected-INT3 floor: (i) MMLU accuracy (capability), (ii) GCG attack ASR (adversarial robustness), (iii) XSTest safety score (over-refusal balance). On weakly-aligned models the scheme can be net-harmful: GCG ASR under protection exceeds that of unprotected uniform quantization.*

The proposition asserts that refusal-template completion is a *separate channel* from the safety properties measured by capability, adversarial, and over-refusal probes. We establish it by exhibiting concrete protection schemes satisfying each clause (§8, four exhibits); a single satisfying instance suffices for an existence claim.

5 Experimental Setup

Models. We evaluate on a 5-model panel spanning the alignment spectrum: **Llama-3.1-8B-Instruct** (Llama Team, 2024) (Meta, multi-round safety RLHF, strongly aligned), **Qwen3-8B** (Qwen Team, 2025) (Alibaba, RLHF, strongly aligned), **GLM-4-9B-0414** (Z.ai, 2025) (Zhipu, RLHF, strongly aligned), **Ministral-8B-Instruct-2410** (Mistral AI, 2024) (Mistral, no published safety RLHF, weakly aligned), **DeepSeek-R1-Distill-Llama-8B** (DeepSeek-AI, 2025) (DeepSeek, reasoning-distilled *without* safety alignment, cf. Jeung et al. (2025), Zhang et al. (2025b) which document this gap).

Quantization. Symmetric group-wise simulated quant-dequant, group size 128, primarily at 3-bit (the setting where safety collapse is observable); 4-bit ablation in §9.

Datasets. AdvBench (Zou et al., 2023) (520 prompts) for safety-loss gradient computation; HarmBench (Mazeika et al., 2024) (400 prompts)

for refusal rate; JBB-Behaviors GCG-attack artifacts (Chao et al., 2024) (100 transferred from Llama-2-7B-chat); XSTest (Röttger et al., 2024) (a balanced subset of 100 safe + 100 unsafe prompts); WikiText-2 (Merity et al., 2017) for perplexity; MMLU-lite (Hendrycks et al., 2021) (200 questions, lightweight capability evidence); PAIR transfer artifacts (Chao et al., 2025) from Jailbreak-Bench for attack-family generalization (App. L).

Attack metrics. **GCG attack ASR** is reported on two metrics throughout. The *heuristic non-refusal rate* is the fraction of GCG-suffixed prompts without a refusal, a standard transfer-attack proxy (the refusal rate uses the same detector). The *classifier-judged harmful-response rate* runs the HarmBench-Llama-2-13b classifier (Mazeika et al., 2024) on the same completions. The two dissociate at 3 bits, where degraded models produce compliant but incoherent output (§8); the classifier column carries the harm claims, and Table 1 reports both.

Scoring rules. We test SSMP $\sum_j |g_j \delta w_j|$, SNIP $\sum_j |g_j w_j|$, magnitude $|w|$, and gradient-only/Fisher $\|g\|_2^2$ at protection budgets $\beta \in \{4\%, 8\%, 60\%\}$. The 60% budget mirrors CWP’s published recipe; Fisher@60% is the closest no-retraining stand-in for CWP’s Fisher-family score.

Statistical methods. Generation is deterministic (temperature 0), so per-condition refusal is exact; the remaining variance is over the finite evaluation-prompt set. We therefore report paired-over-prompts bootstrap confidence intervals (10k resamples, same resampled prompt indices for both scorers) for pairwise scorer differences (Table 19), with matching bootstrap CIs for GCG ASR and XSTest and Wilson intervals for MMLU in App. R.

Main results. Table 1 reports the complete cross-model panel at 3-bit on the four primary metrics, with the GCG column carrying both the heuristic and the classifier judgment. SSMP and SNIP (the two-factor pair) track each other within a few pp on every metric on every model, the direct Theorem 2 test; pure magnitude, a single-factor escape, can differ substantially (e.g. 17pp from SSMP on Qwen3 refusal). The table also previews four findings detailed below: the heuristic-vs-classifier dissociation on coherence-degraded models (Qwen3 INT3: 0.91 vs. 0.03), Fisher’s escape on Llama under GCG (§7.2), the net-harmful protection effect

on Ministral (§8), and multi-dimensional decoupling at Llama Fisher@60% (§8). The four exhibits are extracted in Table 12.

6 Validating the Drift Bound (Theorem 1)

Figure 2 reports the bound across the 5 models, on the denser 6-point bit-width grid for Llama-3.1 and Qwen3. Three observations. (i) The bound is a strict upper envelope: predicted $\geq$ measured at every point. (ii) It orders drift severity: measured $|\Delta L_{\text{safe}}|$ stays below 0.3 at every bit-width ≥ 3 and jumps to 12–18 at 2-bit, with pooled rank correlation 0.834 over all 24 points. The sub-ordering among the near-zero ≥ 3 -bit points is evaluation noise; the floor-vs-catastrophe ordering is what the theorem asserts and what protection decisions depend on. (iii) Magnitude-tightness varies as Theorem 2 predicts (R^2 from 0.90 on Llama-3.1 and Qwen3 down to 0.01 on DeepSeek, panel titles): per-group constants set the bound’s scale per model while the ordering is preserved.

7 Validating Structural Collapse (Theorem 2)

We validate Theorem 2 at two levels: the *mechanism* (which blocks each scorer selects, §7.1) and the *downstream outcome* (the safety metrics those selections produce, §7.2).

7.1 Direct test: protect-set agreement

Corollary 1 makes a prediction about *block selection*, testable without reference to any downstream metric: the two gradient-weighted scorers (SSMP, SNIP) should rank and select near-identical blocks, while the single-factor scorers diverge. Table 2 confirms this directly from the raw block scores. Over *all* blocks, SSMP and SNIP are near-perfectly rank-correlated (Spearman 0.997–1.000); their top-8% protect sets overlap with Jaccard 0.89–0.97 (random baseline 0.042): the same weights both globally and where it matters. Fisher (magnitude dropped) overlaps the SSMP/SNIP pair far less (Jaccard 0.19–0.55), and the reason is structural: the escape-clause correlation $\text{corr}(\text{Fisher}, \text{magnitude})$ over the top-8% blocks is *negative on all five models* (−0.33 to −0.68), so the highest-gradient blocks are systematically not the highest-magnitude ones. Pure magnitude, which drops the gradient factor entirely, diverges most; the full pairwise Jaccard heatmap (Fig. 5, App. O) and per-scorer Spearman break-

Method	Llama-3.1				Qwen3				GLM-4				Ministral				DeepSeek			
	R $\uparrow$	G _{h/c} $\downarrow$	X $\uparrow$	M $\uparrow$	R $\uparrow$	G _{h/c} $\downarrow$	X $\uparrow$	M $\uparrow$	R $\uparrow$	G _{h/c} $\downarrow$	X $\uparrow$	M $\uparrow$	R $\uparrow$	G _{h/c} $\downarrow$	X $\uparrow$	M $\uparrow$	R $\uparrow$	G _{h/c} $\downarrow$	X $\uparrow$	M $\uparrow$
FP16	0.68	0.06/0.04	0.93	0.45	0.58	0.21/0.08	0.85	0.62	0.57	0.12/0.07	0.93	0.52	0.27	0.66/0.23	0.62	0.30	0.00	1.00/0.15	0.00	0.25
INT3 (uniform 3-bit)	0.37	0.53/0.28	0.62	0.24	0.10	0.91/0.03	0.29	0.27	0.55	0.12/0.03	0.91	0.41	0.16	0.87/0.16	0.50	0.28	0.00	1.00/0.05	0.00	0.27
SSMP @ 8%	0.42	0.59/0.25	0.32	0.27	0.31	0.88/0.03	0.57	0.32	0.60	0.10/0.04	0.90	0.41	0.17	0.94/0.22	0.57	0.24	0.00	1.00/0.04	0.00	0.23
SNIP @ 8%	0.44	0.59/0.27	0.35	0.27	0.33	0.89/0.07	0.48	0.33	0.62	0.12/0.04	0.91	0.36	0.18	0.93/0.25	0.56	0.23	0.00	1.00/0.06	0.00	0.27
Magnitude @ 8%	0.47	0.50/0.24	0.37	0.26	0.14	0.99/0.03	0.46	0.29	0.54	0.13/0.05	0.92	0.42	0.15	0.89/0.22	0.47	0.29	0.00	1.00/0.04	0.00	0.26
Fisher @ 4%	0.55	0.62/0.27	0.57	0.30	0.35	0.92/0.00	0.45	0.33	0.58	0.09/0.04	0.91	0.36	0.15	0.91/0.34	0.50	0.29	0.00	1.00/0.10	0.00	0.30
Fisher @ 8%	0.51	0.37/0.15	0.54	0.34	0.46	0.87/0.04	0.42	0.34	0.57	0.09/0.04	0.91	0.40	0.19	0.93/0.32	0.49	0.28	0.00	1.00/0.07	0.00	0.26
Fisher @ 60%	0.74	0.44/0.18	0.85	0.33	0.66	0.52/0.05	0.54	0.36	0.55	0.14/0.09	0.93	0.37	0.24	0.85/0.26	0.47	0.28	0.00	1.00/0.14	0.00	0.21

Table 1: **Main results** at 3-bit symmetric group-wise PTQ across 5 open-weight models. Columns per model: **R** = HarmBench refusal rate ($\uparrow$ safer); **G_{h/c}** = JBB-GCG attack success rate, reported as heuristic non-refusal rate / classifier-judged harmful-response rate (HarmBench-Llama-2-13b; $\downarrow$ safer for both); **X** = XSTest safety score ($\uparrow$ better balance), **M** = MMLU-lite accuracy ($\uparrow$ more coherent). Rows: FP16 reference, unprotected INT3, the two gradient-weighted scorers (SSMP, SNIP) and pure magnitude at 8% budget, and the Fisher family at 4%, 8%, and 60% budgets. The two GCG numbers dissociate on coherence-degraded models (Qwen3 INT3: 0.91 heuristic vs. 0.03 classifier): the heuristic counts incoherent non-refusals as attack successes, so the classifier column carries the harm claim (§8). Refusal restoration is decoupled from adversarial robustness on both metrics (Proposition 3); §7–§8 analyze the block-selection and outcome structure.

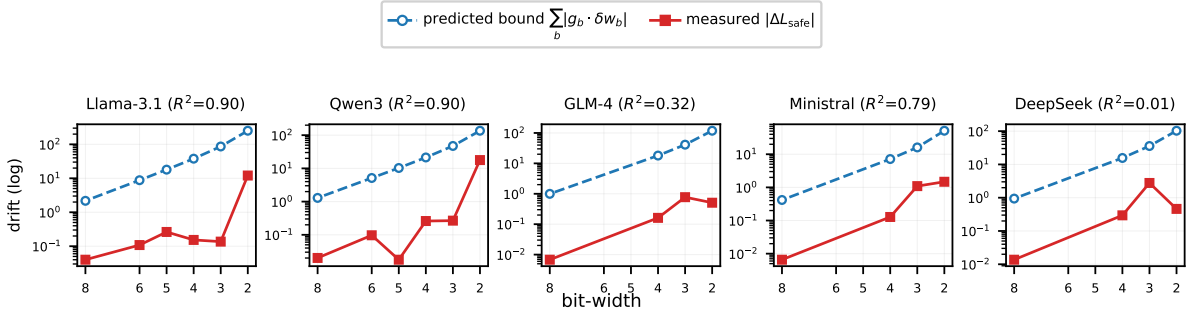


Figure 2: **Theorem 1 validation (one panel per model)**. Predicted upper bound $\sum_b |g_b \cdot \delta w_b|$ (dashed) and measured $|\Delta L_{\text{safe}}|$ (solid) per bit-width, log scale; Llama-3.1 and Qwen3 use the denser $\{8, 6, 5, 4, 3, 2\}$ grid, the others $\{8, 4, 3, 2\}$. This figure supports the drift bound (Theorem 1); the structural-collapse evidence for Theorem 2 is Figure 3 and Table 2. The bound is a strict upper envelope that orders drift severity: pooled over all 24 (model $\times$ bit-width) points, Spearman $\rho = 0.834$, and the 2-bit catastrophe is separated from the benign ≥ 3 -bit range by $40\text{--}70\times$.

down (Table 18, App. O) make the 2×2 factor structure explicit. This confirms structural collapse at the level of *which weights are protected*, prior to any evaluation; whether the divergent Fisher selection changes the downstream metric depends on the model category (§7.2). Apps. C–F extend the equivalence to the full compound family, other block sizes, other quantizer structures, and 14B.

7.2 Downstream outcome: cluster + escape

Outcome clustering. For each (model, budget) cell, we compute the spread of heuristic GCG attack ASR across the three non-Fisher scorers: it is at most 0.11 in every cell (Table 9, App. J); Figure 3 visualizes the result.

Fisher escape. Theorem 2’s secondary prediction is that scorers *outside* the two-factor gradient–

magnitude family (Fisher g^2 and, in principle, pure magnitude) can produce a distinct ranking. Fisher does so on *Llama* at $\beta=8\%$ (Table 9), and the escape survives the classifier metric: Fisher’s classifier-judged ASR there is 0.15 vs. the two-factor cluster’s 0.24–0.27 (Table 1). On the other models its value falls within the cluster, either because the protected model is at a ceiling (Qwen3, Ministral, DeepSeek under GCG) or the unprotected model is already robust (GLM-4): the predicted category-dependent manifestation analyzed in §9.

Statistical significance. Table 19 (App. R) reports paired-over-prompts bootstrap CIs for the SSMP-vs-other refusal-rate differences at $\beta=8\%$. The SSMP–SNIP CI contains 0 on 4 of 5 models (Ministral’s gap is 1.2pp), confirming the

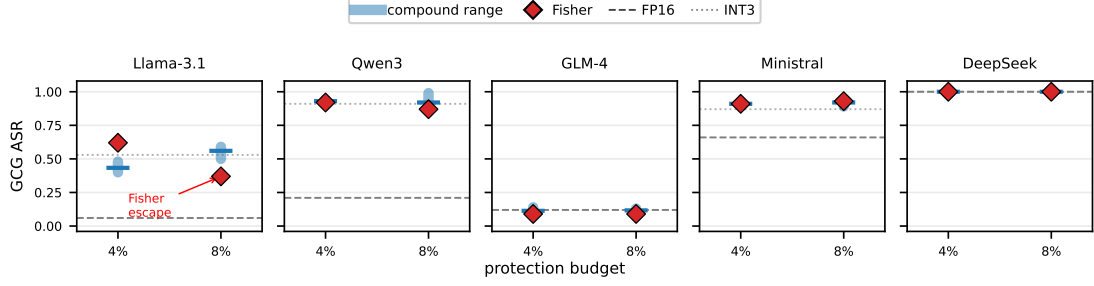


Figure 3: **Theorem 2: cluster + Fisher escape (heuristic GCG ASR), one panel per model.** x -axis: protection budget (4%, 8%). Blue bar = range of the three non-Fisher scorers (SSMP, SNIP, magnitude); red diamond = Fisher; dashed/dotted = FP16/INT3 references. Non-Fisher scorers cluster tightly everywhere; Fisher escapes the cluster only on Llama-3.1 @8% (0.37 vs. cluster mean 0.56). The same cluster-plus-escape structure holds under the classifier metric (Table 1).

Model	SSMP $\equiv$ SNIP		Fisher	Escape
	ρ_{all}	Jac _{8%}	Jac _{8%}	corr
Llama-3.1	0.998	0.91	0.27	-0.41
Qwen3	0.999	0.89	0.19	-0.46
GLM-4	0.999	0.97	0.55	-0.33
Ministral	1.000	0.97	0.33	-0.68
DeepSeek	0.997	0.91	0.40	-0.54

Table 2: **Direct test of Theorem 2 (block selection).** SSMP and SNIP are the two gradient-weighted scorers. ρ_{all} : Spearman correlation of their block scores over *all* blocks (from raw scores); Jac_{8%}: Jaccard overlap of their top-8% protect sets (random baseline 0.04). Both confirm SSMP and SNIP are near-interchangeable. **Fisher** Jac_{8%} is the mean overlap of Fisher’s protect set with the SSMP/SNIP pair — far lower, i.e. Fisher escapes. **Escape** corr is Spearman(Fisher, magnitude) on the top-8% blocks: negative on all models, explaining why Fisher’s magnitude-free ranking selects different blocks.

two-factor pair is statistically indistinguishable; SSMP–Fisher is significant exactly on the models between the floor and the ceiling (Llama, Qwen3, GLM-4), the escape pattern, and the same holds on GCG and XSTest (App. R).

8 Validating Decoupling (Proposition 3)

Exhibit criteria. For any metric m we define the recovery ratio $\rho = (m_{\text{cond}} - m_{\text{INT3}})/(m_{\text{FP16}} - m_{\text{INT3}})$, with fixed labels applied uniformly to every cell: RESTORED $\rho \geq 1/3$; PARTIAL $0.10 \leq \rho < 1/3$; FLOOR $|\rho| < 0.10$; NET-HARMFUL $\rho \leq -0.10$; N/A when the $|\text{FP16} - \text{INT3}|$ gap is degenerate (< 0.05 ; < 1.0 for PPL). App. N labels all cells; the exhibits below follow systematically from this single definition.

Figure 1 (§1) gives the 2D decoupling map; we

extract four exhibits, each demonstrating a different decoupled axis (Table 12; detail tables in App. K). **(A) Capability.** Magnitude@4% on Llama-3.1 (3-bit) restores refusal to 0.48 ($\rho = +0.36$, RESTORED) while MMLU falls to chance (0.245, Wilson 95% [0.19, 0.31]; FLOOR) and PPL triples to 27.3: refusal-template completion survives as capability collapses, independent of any attack metric. **(B) Behavioral metrics.** On Qwen3, Fisher@8% recovers 73% of the refusal-rate gap (RESTORED) but only 6% of the heuristic GCG-ASR gap (FLOOR; recovery gap $+0.68$ [$+0.55, +0.82$]); across all four scorers refusal recovery exceeds attack-metric recovery by 6–12 $\times$ (Table 13). Qwen3’s 3-bit “attack successes” are incoherent non-refusals rather than harmful content (classifier ASR 0.03 at INT3 vs. 0.08 at FP16): both behavioral metrics move independently of classifier-judged harm. **(C) Net-harmful protection.** On weakly-aligned Ministral, Fisher protection *raises classifier-judged harm* above the unprotected floor: 0.34 at 4% and 0.32 at 8% vs. INT3’s 0.16 ($+0.18$ [$+0.06, +0.30$] and $+0.16$ [$+0.05, +0.27$], CIs exclude 0), above even FP16’s 0.23; the heuristic agrees in direction (Table 14). Fisher is specifically the net-harmful scorer (the SSMP, SNIP, and magnitude deltas are not significant): protecting mis-identified blocks injects perturbation without recovering safety, because the weak safety mechanism produces no sharp gradients to localize. The net-harmful *direction* is GCG-specific (neutral under PAIR, App. L). **(D) Multi-dimensional dissociation.** On Llama, Fisher@60% restores refusal (0.745, above FP16’s 0.677) and XSTest balance (0.85 vs. FP16’s 0.93), while classifier GCG ASR stays at 0.18 vs. FP16’s

0.04 (4.5 $\times$; heuristic 0.44 vs. 0.06). Together the exhibits decouple refusal-template completion from every safety property we measure (A: capability; B, C, D: classifier-judged harm; C: direction of effect; D: XSTest–GCG dissociation), each independently establishing Proposition 3.

Activation-level evidence. The decoupling is visible below behavior. On Llama-3.1, the harmful-vs-harmless separation along the refusal direction (Arditi et al., 2024) in the residual stream is essentially unchanged across FP16 (37.5), INT3 (37.0), and Fisher@8% (37.0), while refusal behavior swings 0.68 $\rightarrow$ 0.37 $\rightarrow$ 0.51 (Table 16, App. M): quantization’s safety failure, and protection’s partial recovery, are downstream of an intact refusal representation.

Attack-family generalization. The decoupling reproduces under PAIR (Chao et al., 2025), a structurally different semantic jailbreak: at INT3 the heuristic over-states classifier-judged harm by +0.43 to +0.93 across models, and protection again leaves classifier-judged ASR at the unprotected level (App. L).

9 Cross-Model Analysis

Theorems 1–2 and Proposition 3 hold *uniformly* across all five models: the drift bound upper-envelopes the measured drift everywhere (Fig. 2), SSMP $\equiv$ SNIP everywhere (Table 2), and decoupling appears on every model with nonzero refusal (§8). The four-category classification (Table 10, App. J) organizes a separate question: whether the theory’s predictions produce a *measurable downstream effect*, which requires the model to sit strictly between the safety floor (unprotected INT3) and ceiling (FP16). That residual variation is predictable from one FP16 baseline metric: *the heuristic GCG non-refusal rate*, a behavioral measure of how much refusal behavior the model has to lose (classifier-judged harm instead concentrates on the models that stay coherent at 3 bits, §8).

Why the category determines theorem visibility. T2’s universal cluster prediction holds in every category. T2’s secondary “Fisher escape” prediction requires the cluster to lie strictly between the floor (unprotected INT3) and the ceiling (FP16). Only Llama-3.1 (strongly aligned and quantization-fragile) meets this condition, so the escape is visible there (heuristic: cluster at 0.56, Fisher at 0.37, FP16 at 0.06; classifier: 0.24–0.27, 0.15, 0.04). On

Qwen3 the cluster sits at the heuristic ceiling, on GLM-4 at the FP16 baseline, and on Ministral and DeepSeek at the floor.

Bit-width effects. At 4-bit, none of the five models exhibits the catastrophic collapse seen at 3-bit (refusal drop at 4-bit vs. 3-bit: Qwen3 -8.5 vs. -48.0 pp, Llama $+4.2$ vs. -31.0 , GLM-4 $+0.3$ vs. -2.0 , Ministral -1.8 vs. -11.0 ; the DeepSeek comparison is degenerate with FP16 refusal 0; Table 11, App. J). Production deployments operating at 4-bit therefore sit in a setting where all protection schemes and unprotected uniform quantization yield approximately equivalent behavior. The relevant setting for safety-aware quantization is the 3-bit boundary, where the theorems make sharp predictions.

10 Discussion and Conclusion

This paper’s central message is a caution for safety-aware quantization. At ≤ 3 bits, the refusal behavior that small-budget weight protection recovers is decoupled from the safety it is meant to preserve (Proposition 3): from capability, from classifier-judged harm under two attack families, from over-refusal balance, and even from the refusal representation. A different compound scorer leaves this unchanged, because the entire gradient-weighted family $\alpha|g \cdot w| + \beta|g \cdot \delta w|$ selects the same blocks (Theorem 2), invariant to block size, quantizer structure, and scale. The drift bound (Theorem 1) orders drift severity and locates the 3-bit boundary; a four-category classification of models organizes the panel.

What the theorems tell practitioners. Three takeaways for mixed-precision safety protection at ≤ 3 bits: **(1)** the choice among compound gradient-weighted scorers is immaterial for block ranking; only dropping a factor (magnitude, Fisher) changes the selection. **(2)** verify claimed safety recovery with classifier-judged adversarial evaluation: refusal can recover while capability and robustness stay at the floor, and at 3 bits heuristic metrics conflate incoherence with harm. **(3)** on weakly-aligned models (FP16 heuristic GCG ASR $> 50\%$) protection can be net-harmful; prefer ≥ 4 bits.

Limitations

(i) Model scale: we study the 8–9B instruction-tuned class, where 3-bit safety degradation is most acute and most deployment-relevant. The per-block

analysis of Theorem 2 is scale-independent, its concentration tightens with the hidden dimension, and we verify the block-selection equivalence directly at 14B (App. F); a full downstream safety sweep at 70B and above exceeds our 48 GB hardware and is natural future work. (ii) Simulated quantization: the main panel uses the arithmetic of NF4 group-wise quant without storage compression, for reproducibility, and we verify score-ranking equivalence to bnb-NF4 on a sample; the collapse itself is additionally confirmed under NF4, asymmetric, per-channel, and error-feedback quantizer structures (App. E), though the downstream safety metrics are measured under the symmetric group-wise quantizer only. (iii) Baseline scope: we isolate the *scoring rule*. We additionally run the full Q-resafe (SNIP+LoRA-DPO) pipeline on one model, which lands in the same recovery range as small-budget scoring, consistent with the analysis (App. S); CWP’s Fisher-family scoring is represented by our Fisher@60% ablation, and its exact Fisher-difference recipe is not re-run. (iv) Adversarial robustness is measured under transferred GCG and PAIR artifacts, judged by both a heuristic non-refusal detector and the HarmBench classifier; adaptive (per-model optimized) attacks are not run, and the activation analysis uses a single-direction refusal proxy (§5, App. M). (v) Greedy decoding is deterministic, so we report paired-over-prompts bootstrap CIs rather than a full multi-seed sweep; MMLU-lite (200 questions) is lightweight capability evidence and is phrased as such.

Ethics Statement

This work uses public safety benchmarks that intentionally contain harmful or adversarial prompts; we collect no new data, release no new harmful content, and report only aggregate metrics. The findings are diagnostic: they show that refusal-rate-only evaluation over-states the safety of aggressively quantized models. The population at risk from that over-statement is users of compressed models in resource-constrained deployments (on-device assistants, edge serving, low-cost APIs), where ≤ 3 -bit quantization is most attractive and where a model that emits refusal templates without mechanistic robustness gives operators false assurance. Publishing these failure modes could conceivably be used to rationalize shipping an unsafely quantized model, but we judge the greater risk to be the status quo, in which superficial refusal re-

covery passes for safety. We therefore recommend classifier-judged adversarial evaluation before deploying any model quantized to ≤ 3 bits, and avoiding ≤ 3 -bit quantization of weakly-aligned models entirely.

Acknowledgments

We thank the three anonymous ARR reviewers and the action editor, whose requests shaped this version’s classifier-judged adversarial evaluation, second attack family, quantizer-structure analysis, and uniform exhibit criteria. This work is supported in part by the National Science Foundation under Award #2530909 and by NEUTC.

References

- Muhammad Alif Al Hakim, Alfian Farizki Wicaksono, and Fajri Koto. 2026. [Preserving fairness and safety in quantized LLMs through critical weight protection](#). In *Findings of the Association for Computational Linguistics: ACL 2026*, pages 19831–19855, San Diego, California, United States. Association for Computational Linguistics.
- Andy Arditi, Oscar Obeso, Aaqib Syed, Daniel Paleka, Nina Panickssery, Wes Gurnee, and Neel Nanda. 2024. [Refusal in language models is mediated by a single direction](#). *Preprint*, arXiv:2406.11717.
- Lang Cao. 2024. [Learn to refuse: Making large language models more controllable and reliable through knowledge scope limitation and refusal mechanism](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 3628–3646. Association for Computational Linguistics.
- Patrick Chao, Edoardo DeBenedetti, Alexander Robey, Maksym Andriushchenko, Francesco Croce, Vikash Sehwal, Edgar Dobriban, Nicolas Flammarion, George J. Pappas, Florian Tramèr, Hamed Hassani, and Eric Wong. 2024. [JailbreakBench: An open robustness benchmark for jailbreaking large language models](#). In *Advances in Neural Information Processing Systems*, volume 37, pages 55005–55029. Curran Associates, Inc.
- Patrick Chao, Alexander Robey, Edgar Dobriban, Hamed Hassani, George J. Pappas, and Eric Wong. 2025. [Jailbreaking black box large language models in twenty queries](#). In *2025 IEEE Conference on Secure and Trustworthy Machine Learning*, pages 23–42. IEEE.
- Kejia Chen, Jiawen Zhang, Jiacong Hu, Yu Wang, Jian Lou, Zunlei Feng, and Mingli Song. 2025. [Assessing safety risks and quantization-aware safety patching for quantized large language models](#). In *Proceedings of the 42nd International Conference on Machine*

- Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 9728–9746. PMLR.
- DeepSeek-AI. 2025. [DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning](#). Preprint, arXiv:2501.12948.
- Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. [QLoRA: Efficient fine-tuning of quantized LLMs](#). In *Advances in Neural Information Processing Systems*, volume 36, pages 10088–10115. Curran Associates, Inc.
- Elias Frantar, Saleh Ashkboos, Torsten Hoefler, and Dan Alistarh. 2023. [GPTQ: Accurate post-training quantization for generative pre-trained transformers](#). In *The Eleventh International Conference on Learning Representations*.
- Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. [Measuring massive multitask language understanding](#). In *International Conference on Learning Representations*.
- Fabian Hildebrandt, Andreas K. Maier, Patrick Krauss, and Achim Schilling. 2025. [Refusal behavior in large language models: A nonlinear perspective](#). In *International Joint Conference on Neural Networks*, pages 1–7. IEEE.
- Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. [LoRA: Low-rank adaptation of large language models](#). In *10th International Conference on Learning Representations*. OpenReview.net.
- Wei Huang, Haotong Qin, Yangdong Liu, Yawei Li, Qinsuo Liu, Xianglong Liu, Luca Benini, Michele Magno, Shiming Zhang, and Xiaojuan Qi. 2025. [SLiM-LLM: Saliency-driven mixed-precision quantization for large language models](#). In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 25672–25692. PMLR.
- Wonje Jeung, Sangyeon Yoon, Minsuk Kahng, and Albert No. 2025. [SAFEPath: Preventing harmful reasoning in chain-of-thought via early alignment](#). In *Advances in Neural Information Processing Systems*, volume 38, pages 110480–110509. Curran Associates, Inc.
- Namhoon Lee, Thalaiyasingam Ajanthan, and Philip H. S. Torr. 2019. [SNIP: Single-shot network pruning based on connection sensitivity](#). In *International Conference on Learning Representations*.
- Shiyao Li, Xuefei Ning, Ke Hong, Tengxuan Liu, Lun-ing Wang, Xiuhong Li, Kai Zhong, Guohao Dai, Huazhong Yang, and Yu Wang. 2023. [LLM-MQ: Mixed-precision quantization for efficient LLM deployment](#). In *Third Workshop on Efficient Natural Language and Speech Processing*. ENLSP-III at NeurIPS 2023.
- Ji Lin, Jiaming Tang, Haotian Tang, Shang Yang, Wei-Ming Chen, Wei-Chen Wang, Guangxuan Xiao, Xingyu Dang, Chuang Gan, and Song Han. 2024. [AWQ: Activation-aware weight quantization for on-device LLM compression and acceleration](#). In *Proceedings of Machine Learning and Systems*, volume 6, pages 87–100.
- Llama Team. 2024. [The Llama 3 herd of models](#). Preprint, arXiv:2407.21783.
- Mantas Mazeika, Long Phan, Xuwang Yin, Andy Zou, Zifan Wang, Norman Mu, Elham Sakhaee, Nathaniel Li, Steven Basart, Bo Li, David A. Forsyth, and Dan Hendrycks. 2024. [HarmBench: A standardized evaluation framework for automated red teaming and robust refusal](#). In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 35181–35224. PMLR.
- Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. 2017. [Pointer sentinel mixture models](#). In *5th International Conference on Learning Representations*. OpenReview.net.
- Mistral AI. 2024. [Minstral 8B Instruct 2410](#). Model card.
- Sindhu Padakandla, Sadbhavana Babar, Rathod Darshan D, and Manohar Kaul. 2025. [SafeQuant: LLM safety analysis via quantized gradient inspection](#). In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2522–2536, Albuquerque, New Mexico. Association for Computational Linguistics.
- Hoang Pham, The-Anh Ta, and Long Tran-Thanh. 2026. [Pruning at initialisation through the lens of graphon limit: Convergence, expressivity, and generalisation](#). Preprint, arXiv:2602.06675. Accepted to the 43rd International Conference on Machine Learning.
- Qwen Team. 2025. [Qwen3 technical report](#). Preprint, arXiv:2505.09388.
- Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D. Manning, Stefano Ermon, and Chelsea Finn. 2023. [Direct preference optimization: Your language model is secretly a reward model](#). In *Advances in Neural Information Processing Systems*, volume 36, pages 53728–53741. Curran Associates, Inc.
- Paul Röttger, Hannah Kirk, Bertie Vidgen, Giuseppe Attanasio, Federico Bianchi, and Dirk Hovy. 2024. [XSTest: A test suite for identifying exaggerated safety behaviours in large language models](#). In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5377–5400, Mexico City, Mexico. Association for Computational Linguistics.

- Jianchang Su, Masoud Rahimi Jafari, Yifan Zhang, and Wei Zhang. 2024. [FastMatch: Enhancing data pipeline efficiency for accelerated distributed training](#). In *2024 IEEE 42nd International Conference on Computer Design*, pages 239–246. IEEE.
- Jianchang Su, Wenhui Zhang, Yifan Zhang, Kexin Chu, Hao Guo, Youyou Lu, and Wei Zhang. 2026. [TrustWeave: Integrity measurement and attestation for multi-cloud LLMs](#). In *Proceedings of the 21st European Conference on Computer Systems*, pages 641–656. ACM.
- Hidenori Tanaka, Daniel Kunin, Daniel L. K. Yamins, and Surya Ganguli. 2020. [Pruning neural networks without any data by iteratively conserving synaptic flow](#). In *Advances in Neural Information Processing Systems*, volume 33.
- Chaoqi Wang, Guodong Zhang, and Roger Grosse. 2020. [Picking winning tickets before training by preserving gradient flow](#). In *8th International Conference on Learning Representations*. OpenReview.net.
- Xinpeng Wang, Mingyang Wang, Yihong Liu, Hinrich Schütze, and Barbara Plank. 2025. [Refusal direction is universal across safety-aligned languages](#). In *Advances in Neural Information Processing Systems*, volume 38, pages 32380–32423. Curran Associates, Inc.
- Sunghyun Wee, Suyoung Kim, Hyeonjin Kim, Kyomin Hwang, and Nojun Kwak. 2025. [Safety-preserving PTQ via contrastive alignment loss](#). *Preprint*, arXiv:2511.07842.
- Bernard Widrow, István Kollár, and Ming-Chang Liu. 1996. [Statistical theory of quantization](#). *IEEE Transactions on Instrumentation and Measurement*, 45(2):353–361.
- Tom Wollschläger, Jannes Elstner, Simon Geisler, Vincent Cohen-Addad, Stephan Günnemann, and Johannes Gasteiger. 2025. [The geometry of refusal in large language models: Concept cones and representational independence](#). In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 66945–66970. PMLR.
- An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, and 22 others. 2024. [Qwen2.5 technical report](#). *Preprint*, arXiv:2412.15115.
- Zhen Yang, Sizai Hou, Kaiwen Zheng, Yaofang Liu, Liang He, Yixuan Chen, and Kangning Cui. 2026. [Target-aware calibration data selection for preserving uncertainty in quantized language models](#). *Preprint*, arXiv:2608.21019.
- Z.ai. 2025. [GLM-4-9B-0414](#). Model card.
- Ruyi Zhang, Heng Gao, Songlei Jian, Yusong Tan, and Haifang Zhou. 2026a. [BadLLM-TG: A backdoor defender powered by LLM trigger generator](#). In *ICASSP 2026—2026 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 13862–13866. IEEE.
- Ruyi Zhang, Songlei Jian, Yusong Tan, Heng Gao, Haifang Zhou, and Kai Lu. 2025a. [BadWindtunnel: Defending backdoor in high-noise simulated training with confidence variance](#). In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 9259–9273, Vienna, Austria. Association for Computational Linguistics.
- Ruyi Zhang, Hongzuo Xu, Songlei Jian, Yusong Tan, Haifang Zhou, and Rulin Xu. 2026b. [Angel or devil: Discriminating hard samples and anomaly contaminations for unsupervised time series anomaly detection](#). *Neural Networks*, 197:108532.
- Yichi Zhang, Zihao Zeng, Dongbai Li, Yao Huang, Zhi-jie Deng, and Yinpeng Dong. 2025b. [RealSafe-R1: Safety-aligned DeepSeek-R1 without compromising reasoning capability](#). *Preprint*, arXiv:2504.10081.
- Andy Zou, Zifan Wang, Nicholas Carlini, Milad Nasr, J. Zico Kolter, and Matt Fredrikson. 2023. [Universal and transferable adversarial attacks on aligned language models](#). *Preprint*, arXiv:2307.15043.

A Full Proofs

We restate and prove each result. Throughout, a block b has $K = 64$ output rows and d_{in} columns, so $w_b, g_b, \delta w_b \in \mathbb{R}^{Kd_{\text{in}}}$. We write $n = Kd_{\text{in}}$ for the block dimension.

A.1 Proof of Theorem 1 (Drift Bound)

Proof. Let $\Delta L_{\text{safe}} = L_{\text{safe}}(W + \delta W) - L_{\text{safe}}(W)$. Since L_{safe} is twice differentiable, Taylor’s theorem with Lagrange remainder gives

$$\Delta L_{\text{safe}} = \langle \nabla L_{\text{safe}}, \delta W \rangle + \frac{1}{2} \delta W^\top \mathbf{H}(\xi) \delta W$$

for some ξ on the segment $[W, W + \delta W]$, where $\mathbf{H}$ is the Hessian. The quadratic term is $O(\|\delta W\|_2^2)$ provided $\|\mathbf{H}\|_{\text{op}}$ is bounded on the segment. Decomposing the inner product over blocks (the blocks partition the coordinates), $\langle \nabla L_{\text{safe}}, \delta W \rangle = \sum_b \langle g_b, \delta w_b \rangle$, which is the stated first-order expansion.

For the bound, the triangle inequality and per-block Cauchy–Schwarz give

$$|\Delta L_{\text{safe}}| \leq \sum_b |\langle g_b, \delta w_b \rangle| \leq \sum_b \|g_b\|_2 \|\delta w_b\|_2,$$

dropping the $O(\|\delta W\|_2^2)$ remainder which is non-negative-order in the perturbation scale. $\square$

A.2 Proof of Lemma 1 (SSMP score concentration)

Proof. By Def. 1, $|\delta w_j| = s_{\mathcal{G}(j)} |r_j|$ with r_j the rounding residual. Hence the SSMP score (Def. 2) is

$$s_b^{\text{ssmp}} = \sum_{j \in b} |g_j| \delta w_j = \sum_{j \in b} |g_j| s_{\mathcal{G}(j)} |r_j|.$$

Under Assumption 1, $r_j \sim \text{Unif}[-\frac{1}{2}, \frac{1}{2}]$ i.i.d. and independent of g_j , so $\mathbb{E}|r_j| = \frac{1}{4}$ and $\text{Var}(|r_j|) = \frac{1}{12} - \frac{1}{16} = \frac{1}{48}$. With a dominant group scale s_b ,

$$\mathbb{E}[s_b^{\text{ssmp}}] = s_b \mathbb{E}|r| \sum_{j \in b} |g_j| = \frac{s_b}{4} \|g_b\|_1.$$

Crucially s_b^{ssmp} is a sum of $n = K d_{\text{in}}$ non-negative independent terms, so $\text{Var}(s_b^{\text{ssmp}}) = \frac{s_b^2}{48} \|g_b\|_2^2$ and the relative fluctuation is $\frac{\sqrt{\text{Var}}}{\mathbb{E}} = \frac{4}{\sqrt{48}} \frac{\|g_b\|_2}{\|g_b\|_1} = O(1/\sqrt{n})$ for a dense block. By the law of large numbers $s_b^{\text{ssmp}} = \frac{s_b}{4} \|g_b\|_1 (1 + o(1))$, a deterministic limit. (Contrast the signed inner product $g_b \delta w_b = -\sum_j g_j s_{\mathcal{G}(j)} r_j$, which is zero-mean with $\mathbb{E}[(g_b \cdot \delta w_b)^2] = \frac{s_b^2}{12} \|g_b\|_2^2$: normalizing by its RMS yields a CLT limit in *distribution*, not a constant. The scorer used in practice is the concentrating absolute sum.) $\square$

A.3 Proof of Theorem 2 (Structural Collapse)

Proof. Eq. 8 is Lemma 1. We prove the three consequences.

(Collapse.) By Eq. 8, $s_b^{\text{ssmp}} = \frac{s_b}{4} \|g_b\|_1 (1 + o(1))$, so SSMP ranks blocks by $s_b \|g_b\|_1$. SNIP is $s_b^{\text{snip}} = \sum_{j \in b} |g_j| |w_j|$, the gradient magnitudes weighted by the per-coordinate $|w_j|$. Both are the gradient ℓ_1 mass weighted by a weight-magnitude statistic: SSMP uses the single group scale $s_b = \max_{\mathcal{G}} |w|/q_{\max}$ (identical weight for all coordinates in the group), SNIP uses the per-coordinate $|w_j|$. If within-block weight magnitudes are concentrated ($|w_j| = s_b q_{\max} (1 + \eta_j)$ with η_j small for the high- $|g_j|$ coordinates), then $s_b^{\text{snip}} = s_b q_{\max} \|g_b\|_1 (1 + \bar{\eta})$, rank-equivalent to SSMP. We state this as an empirical regularity rather than an assumption: §7 measures the SSMP–SNIP rank correlation directly and finds it near-unity (Spearman 0.997–1.000), so the dispersion term is negligible in practice.

(Magnitude escape.) Pure magnitude $\|w_b\|_1$ retains the weight-magnitude factor but *drops the*

gradient factor entirely; it ranks by weight magnitude alone, diverging from the two-factor class unless $\|g_b\|_1$ is comonotone with $\|w_b\|_1$.

(Fisher escape.) Pure Fisher $\|g_b\|_2^2$ retains only the gradient factor and *drops the per-coordinate weight magnitude*; it diverges from the two-factor class when the highest-gradient blocks are not the highest-weight-magnitude blocks over the top- β set. $\square$

Remark 1 (Escape is a universal mechanism; its *effect* depends on the model category). The escape clause is testable: compute $\text{corr}(\|g_b\|_2, \text{magnitude})$ over the top- β blocks. Empirically this correlation is *negative on all five models* (−0.33 to −0.68; Table 2), so the highest-gradient blocks are systematically *not* the highest-magnitude blocks, and Fisher selects a different protect-set from the two-factor scorers on every model. Whether this different selection *changes the downstream safety metric* depends on the model category: it surfaces as a measurable gain only on Llama@8% and is masked by ceiling and floor effects elsewhere.

B Concentration Diagnostic: Effective Dimension of the Safety Gradient

Lemma 1’s concentration argument depends on the effective spread of the gradient within a block: if the gradient concentrated in a few coordinates, the relative fluctuation could be much larger than the dense-block rate suggests. We therefore compute the participation ratio $n_{\text{eff}}(g) = \|g\|_1^2 / \|g\|_2^2$ of the safety gradient on exactly the blocks the argument is applied to, the selected top-8% protect set, reported as a conservative lower bound that divides out the $B=128$ -batch accumulation (Table 3). Even the single most gradient-concentrated protected block on any model has effective dimension $\geq 1.4 \times 10^3$, so the worst-case Lemma-1 relative fluctuation is $\leq 1.6\%$ ($\leq 0.2\%$ at the median), far below the inter-block score gaps; only $\sim 5\text{--}8\%$ of protected blocks lie within a $2\times$ -worst-fluctuation band of the protect-set cutoff, so the fluctuation cannot materially change the protect set.

Model	n_{eff} (top-8% blocks)			max rel.
	median	p5	min	fluct.
Llama-3.1	3.3e5	7.0e4	2.3e3	0.0121
Qwen3	1.9e5	1.6e4	1.4e3	0.0155
GLM-4	2.2e5	3.7e4	1.9e3	0.0131
Ministral	2.3e5	1.2e4	2.0e3	0.0128
DeepSeek	3.2e5	1.8e4	2.5e3	0.0117

Table 3: **Effective dimension of the safety gradient on the selected blocks.** Conservative lower bound $n_{\text{eff}} \geq \|g\|_1^2 / (B \|g\|_2^2)$ over the $B=128$ -batch accumulation; the Lemma 1 relative fluctuation is $\leq 0.577/\sqrt{n_{\text{eff}}}$ (last column, worst case over the protect set). Even the most gradient-concentrated protected block has effective dimension $\gtrsim 10^3$, so the concentration step of Theorem 2 incurs $\lesssim 2\%$ worst-case fluctuation.

C The Compound-Scorer Family Collapses as a Continuum

Table 4 extends the two-scorer equivalence of §7.1 to the entire compound family $s(\alpha, \beta) = \alpha|g \cdot w| + \beta|g \cdot \delta w|$: the worst-case Spearman correlation with SNIP over the whole (α, β) grid and all five models is 0.9971. The design space of compound safety scorers therefore has effectively two degrees of freedom, keep or drop the gradient factor and keep or drop the magnitude factor; proposing new mixtures cannot produce a new block ranking under group-wise PTQ.

(α, β)	Llama	Qwen3	GLM-4	Minis.	DeepS.
(1, 0) SNIP	1.0000	1.0000	1.0000	1.0000	1.0000
(0.75, 0.25)	0.9999	0.9994	0.9999	1.0000	0.9998
(0.5, 0.5)	0.9995	0.9994	0.9998	1.0000	0.9992
(0.25, 0.75)	0.9990	0.9991	0.9995	0.9999	0.9983
(0, 1) SSMP	0.9983	0.9987	0.9992	0.9998	0.9971
(1, 1)	0.9995	0.9994	0.9998	1.0000	0.9992
(2, 1)	0.9998	0.9995	0.9999	1.0000	0.9996
(1, 2)	0.9992	0.9992	0.9996	0.9999	0.9987

Table 4: **The whole compound family collapses to one ranking.** Spearman rank correlation of the combined scorer $s(\alpha, \beta) = \alpha|g \cdot w| + \beta|g \cdot \delta w|$ with pure SNIP over an (α, β) grid (each factor normalized to unit mean). **Worst case over the entire grid and all five models: 0.9971.** Any non-negative combination of the two gradient-weighted factors, including the scoring step of Q-resafe and any future compound scorer of this form, ranks blocks in the same equivalence class; only dropping a factor (pure magnitude or pure Fisher) escapes.

D Block-Size Sensitivity

$K=64$ is the block granularity of the SSMP/Taylor method we analyze and $G=128$ the standard group

Model	K	Spearman(SSMP,SNIP)	Jaccard _{8%}
Llama-3.1-8B	16	0.9978	0.941
	32	0.9981	0.946
	64	0.9983	0.958
	128	0.9985	0.964
Qwen3-8B	16	0.9978	0.948
	32	0.9983	0.948
	64	0.9987	0.955
	128	0.9988	0.952

Table 5: **Block-size (K) robustness of the SSMP $\leftrightarrow$ SNIP equivalence** (group size $G=128$). The rank correlation is ≥ 0.99 and the top-8% protect-set Jaccard ≥ 0.94 at every $K \in \{16, 32, 64, 128\}$, tightening as K grows in line with the $O(1/\sqrt{Kd_{\text{in}}})$ concentration rate of Lemma 1.

size; neither is special to the proofs. Table 5 varies the block size: the SSMP $\leftrightarrow$ SNIP equivalence is essentially invariant to K and tightens monotonically as K grows, matching the $O(1/\sqrt{Kd_{\text{in}}})$ concentration rate, while the SSMP $\leftrightarrow$ magnitude correlation stays near zero or negative at every K (the escape persists). This table and App. F recompute block scores from an independent gradient dump, so their absolute Jaccard values differ slightly from Table 2’s; the equivalence-class structure is identical in both. SNIP, Fisher, and magnitude are G -independent by construction; G affects only SSMP’s δw through the per-group scale, which Theorem 2 shows preserves the block ranking.

E Quantizer-Structure Generality

Theorem 2 is stated for the symmetric group-wise quantizer, but its mechanism only requires the error structure $\delta w \propto (\text{per-group scale}) \times (\text{bounded residual})$. Table 6 recomputes SSMP under different real quantizer structures (reusing one gradient pass) and checks the SSMP $\leftrightarrow$ SNIP collapse under each, including the error-feedback case whose correlated residuals violate Assumption 1: the collapse holds under every structure, so it is a property of the two-factor scorer form rather than of the specific quantizer. Whether safety trades off against a calibration-loss reconstruction objective, which uses different data, is an open question we flag for future work.

Quantizer structure	Llama	Qwen3	GLM-4	Minis.	DeepS.
Sym. group-wise (3-bit)	0.9983	0.9987	0.9992	0.9998	0.9971
NF4 non-uniform (4-bit)	0.9991	0.9990	0.9997	0.9999	0.9987
Asym. affine (3-bit)	0.9987	0.9988	0.9994	0.9998	0.9978
Per-channel (3-bit)	0.9929	0.9962	0.9963	0.9954	0.9870
RTN (i.i.d. residuals)	0.9983	0.9987	–	0.9998	–
Error-feedback (corr.)	0.9982	0.9985	–	0.9998	–

Table 6: **The SSMP $\leftrightarrow$ SNIP collapse under different quantizer structures** (Spearman rank correlation over all blocks; SSMP recomputed with each structure’s δw ; group size 128, NF4 native 64). The collapse holds under every structure tested, including NF4 (the deployed non-uniform 4-bit format), asymmetric affine quantization, per-channel scaling with no grouping (worst case, 0.987), and an error-feedback quantizer of the GPTQ family, whose sequentially propagated rounding errors give *correlated* residuals, the one case where Assumption 1 could break: there it stays within 0.0002 of round-to-nearest. The equivalence is a property of the two-factor scorer form rather than of the specific quantizer.

F Scale: A 14B Test Point

Lemma 1’s relative fluctuation is $O(1/\sqrt{Kd_{\text{in}}})$, so the collapse should hold and tighten as the hidden dimension grows; because the per-block argument is independent of depth, it carries to larger models. Table 7 verifies the block-selection equivalence directly on Qwen2.5-14B (Yang et al., 2024), the largest model our 48 GB hardware can score: the equivalence-class structure is unchanged at 14B. A full downstream safety sweep at this scale and above is future work; the block-selection result is the object of Theorem 2 and is what we measure.

Model	d_{in}	$\rho(\text{SSMP}, \text{SNIP})$	Jac@8%
Qwen3-8B	4096	0.9987	0.955
Llama-3.1-8B	4096	0.9983	0.958
GLM-4-9B	4096	0.9992	0.957
Ministral-8B	4096	0.9998	0.981
DeepSeek-R1-8B	4096	0.9971	0.962
Qwen2.5-14B	5120	0.9993	0.925

Table 7: **The block-selection equivalence holds at 14B.** SSMP $\leftrightarrow$ SNIP Spearman over all blocks and top-8% protect-set Jaccard, $K=64$, $G=128$. Qwen2.5-14B ($d_{\text{in}}=5120$) matches the 8-9B panel, consistent with the $O(1/\sqrt{Kd_{\text{in}}})$ rate of Lemma 1: larger hidden dimension only tightens the concentration. The single-factor escapes persist at 14B (Fisher Jaccard 0.47, magnitude 0.61).

G Notation

Table 8 lists the symbols used throughout the paper. Three granularities recur: scoring and protection operate on *blocks* of $K=64$ output rows; the quantizer sets one scale per *group* of $G=128$ weights within a row; and the protection *budget* β is the fraction of blocks kept in FP16. The quantization error $\delta w_j = -s_g r_j$ factors into the per-group scale and a bounded rounding residual, the structure that drives Theorem 2.

Symbol	Meaning
W_ℓ	weight matrix of Linear module ℓ
b, w_b, g_b	block index, weights, safety gradient
$K = 64$	output rows per block
$G = 128$	quantization group size
$q_{\text{max}} = 2^{b-1} - 1$	max quant level at bit-width b
s_g	group scale $\max_G w /q_{\text{max}}$
$\delta w_j = -s_g r_j$	quantization error
$r_j \in [-\frac{1}{2}, \frac{1}{2}]$	rounding residual
β	protection budget (FP16 fraction)
L_{safe}	refusal-template cross-entropy
κ_b	group dynamic range $s_b q_{\text{max}}/\ w_b\ _2$

Table 8: Notation used throughout.

H Experimental Details

Hardware and software. All experiments run on a single NVIDIA RTX A6000 (48 GB) using PyTorch 2.x and HuggingFace Transformers. Quantization is simulated symmetric group-wise quant-dequant in FP16 (the arithmetic of bnb-NF4 prior to storage compression); we verified score-ranking equivalence to bitsandbytes NF4 on a Qwen3 sample.

Gradient computation. Safety-loss gradients g_b are computed over $\min(400, |\mathcal{D}_{\text{safe}}|)$ AdvBench prompts with the refusal template applied via each model’s chat template, sequence length capped at 1024, batch size 1, in a single backward pass. Block size $K = 64$ output rows.

Generation & evaluation. Completions use greedy decoding (temperature 0), 256 new tokens. Refusal rate uses the heuristic refusal detector of Mazeika et al. (2024); classifier-judged ASR uses the HarmBench Llama-2-13B classifier offline on the stored completions (we verified offline-vs-inline agreement, App. I). GCG uses the 100 JBB white-box artifacts transferred from Llama-2-7B-chat; PAIR uses the JBB PAIR transfer artifacts (Vicuna target, 100 usable prompts), evaluated on 20 conditions (FP16, INT3, Fisher@4%, Fisher@8%

per model). XSTest is evaluated on a balanced subset of 100 safe + 100 unsafe prompts; the safety score is the harmonic mean of safe-compliance and unsafe-refusal rates. PPL is WikiText-2 (Merity et al., 2017) (64 windows, length 1024, stride 512); MMLU-lite is 200 questions sampled from MMLU (Hendrycks et al., 2021).

Budgets and bit-widths. Protection budgets $\beta \in \{4\%, 8\%, 60\%\}$; bit-widths $\{2, 3, 4, 8\}$ for the drift-bound validation, densified to $\{2, 3, 4, 5, 6, 8\}$ on Llama-3.1 and Qwen3 (Fig. 2), 3-bit for the main panel, 4-bit for the ablation (§9).

Seeds. Generation is deterministic at temperature 0, so single-seed point estimates are exact for the decoding step. For variance over the protection-selection and scoring steps we ran seeds $\{0, 1, 2\}$ on the Qwen3 G2 panel (App. Q).

I HarmBench Classifier Validation

We validate the offline HarmBench classifier against the inline classifier on the Qwen3 conditions where both were run (fp16, int4): offline and inline ASR agree to within $\pm 0.5\text{pp}$. We therefore use the single offline classifier pipeline for every classifier-judged ASR in the paper: the full 40-condition GCG panel (Table 1) and the 20-condition PAIR panel (App. L), each judged on the stored greedy completions.

J Per-Condition Results (Full Panel)

Table 1 (main text) reports the four primary metrics per (method, model). Table 9 gives the outcome-cluster spread and Fisher-escape values of §7.2, Table 10 the four-category classification of §9, and Table 11 the 3-bit vs. 4-bit comparison of §9 on the full 5-model panel.

The complete machine-generated per-condition tables, including HarmBench-classifier ASR and WikiText-2 PPL for every (model, scorer, budget) cell, are produced by `scripts/figs/tables_paper.py` from the on-disk run directories and included in the supplementary material (`runs/emnlp_synthesis/`).

Model	Cluster spread		@8% GCG ASR	
	$b=.04$	$b=.08$	Fisher	Cluster
Llama-3.1	0.08	0.09	0.37	0.56
Qwen3	0.02	0.11	0.87	0.92
GLM-4	0.05	0.04	0.09	0.12
Ministral	0.02	0.05	0.93	0.92
DeepSeek	0.00	0.00	1.00	1.00

Table 9: Cluster spread (heuristic GCG ASR range across {SSMP, SNIP, magnitude}) and the Fisher vs. cluster-mean values at $\beta=8\%$ (§7.2). Tight clustering everywhere (spread ≤ 0.11). **Fisher escapes only on Llama @8%** (0.37 vs. 0.56, -19pp ; classifier: 0.15 vs. 0.25), the one model where the cluster sits strictly between the floor and the ceiling.

Model	FP16 GCG (heur.)	Category
Llama-3.1	0.06	Strong, quantization-fragile
GLM-4	0.12	Strong, quantization-robust
Qwen3	0.21	Moderate
Ministral	0.66	Weak
DeepSeek	1.00	Unaligned

Table 10: **Four-category classification of the 5-model panel.** Parameter: FP16 baseline heuristic GCG non-refusal rate, a behavioral measure of how much refusal behavior the model has to lose. The category predicts which manifestations of T2/T3 are observable empirically; ceiling and floor effects in the weak and unaligned categories mask theorem-level escape predictions. Classifier-judged harm instead concentrates on the models that stay coherent at 3-bit (Table 1).

Model	3-bit		4-bit	
	ΔR (pp)	ΔPPL	ΔR (pp)	ΔPPL
Qwen3	-48.0	+18.5	-8.5	+1.0
Llama-3.1	-31.0	+18.4	+4.2	+1.1
GLM-4	-2.0	+11.4	+0.3	+1.8
Ministral	-11.0	+11.4	-1.8	—
DeepSeek	+0.0	+203.5	+0.0	—

Table 11: **3-bit vs. 4-bit safety collapse, full 5-model panel.** Difference of unprotected (quantized) vs. FP16 reference on refusal-rate (percentage points) and WikiText-2 PPL, at 3-bit vs. 4-bit symmetric group-wise PTQ. On every model with nonzero FP16 refusal, the 4-bit drop is far milder than the 3-bit one; the DeepSeek row is degenerate (FP16 refusal 0). The Ministral and DeepSeek 4-bit runs report refusal only (—).

K Decoupling Exhibits: Detailed Tables

Table 12 summarizes the four decoupling exhibits of §8. Table 13 expands Exhibit B: every scorer on Qwen3 recovers refusal-rate faster than heuristic GCG ASR, by 6–12 $\times$, so the gap is a property of small-budget protection rather than of any single scorer. Table 14 expands Exhibit C by budget: on

Ministral, no Fisher budget recovers adversarial robustness on either attack metric.

Exhibit	Observation
A: Capability decoupling	Llama @ magnitude @ 4%: refusal = 0.48 (refusal-template restored, recovery +0.36 RESTORED), MMLU = 0.24 (Wilson 95% [0.19, 0.31], random for 4-way MCQ; FLOOR), PPL = 27.3. Refusal behavior preserved while capability collapses; independent of any attack metric.
B: Behavioral-metric decoupling	Qwen3 @ Fisher @ 8%: refusal-rate recovery = 73% of the FP16–INT3 gap (RESTORED), but heuristic GCG-ASR recovery = 6% (FLOOR); the recovery gap is +0.68 [+0.55, +0.82]. The classifier shows both are behavioral metrics: Qwen3’s 3-bit “attack successes” are incoherent non-refusals rather than harmful content (classifier ASR: INT3 0.03 vs. FP16 0.08).
C: Negative protection	Ministral 3-bit, classifier GCG ASR: unprotected INT3 = 0.16; Fisher@4% = 0.34 and @8% = 0.32 both <i>raise</i> classifier-judged harm above the floor (+0.18 [+0.06, +0.30] and +0.16 [+0.05, +0.27], CIs exclude 0) and above FP16’s 0.23. The heuristic agrees in direction (0.91/0.93 vs. 0.87): protection is <i>net-harmful</i> on the weakly-aligned model.
D: Multi-dimensional dissociation	Llama @ Fisher @ 60%: refusal = 0.74 (above FP16’s 0.68), XSTest = 0.85 ($\approx$ FP16’s 0.93), but classifier GCG ASR = 0.18 vs. FP16’s 0.04 (4.5 $\times$; heuristic 0.44 vs. 0.06). Large protection budget restores refusal and over-refusal balance, while adversarial robustness stays near the unprotected level.

Table 12: **Proposition 3 (Decoupling) — four exhibits across four independent axes**, selected by the fixed recovery-ratio criteria of §8 (RESTORED $\rho \geq 1/3$, FLOOR $|\rho| < 0.1$, NET-HARMFUL $\rho \leq -0.1$; every (model, scorer, budget) cell is labeled in App. N). Refusal-rate restoration against benign HarmBench prompts is decoupled from (A) capability, (B) classifier-judged harm under GCG transfer attacks, (C) the direction of the protection effect, and (D) the simultaneous combination of metrics.

Qwen3-8B (3-bit)	Δ Refusal	Δ GCG-ASR
INT3 baseline	0%	0%
Fisher@4%	+52%	−1%
Fisher@8%	+73%	+6%
SSMP@8%	+43%	+4%
SNIP@8%	+47%	+3%
magnitude@8%	+7%	−11%

Table 13: Exhibit B: refusal-rate recovery vs. heuristic GCG-ASR recovery as % of the FP16-INT3 gap. Across all 4 scoring rules on Qwen3, refusal-rate recovery exceeds attack-metric recovery by 6–12 $\times$. Both are behavioral metrics: classifier-judged harm on Qwen3 is near-zero at 3-bit in every condition (Table 1), so the heuristic “attack successes” are incoherent non-refusals.

Ministral-8B (3-bit)	GCG _h	GCG _c	XSTest
FP16	0.66	0.23	0.62
INT3 (no protection)	0.87	0.16	0.50
Fisher@4%	0.91	0.34	0.50
Fisher@8%	0.93	0.32	0.49
Fisher@60%	0.85	0.26	0.47

Table 14: Exhibit C: on the weakly-aligned Ministral-8B, small-budget protection is net-harmful on both attack metrics (GCG_h = heuristic non-refusal, GCG_c = classifier-judged harm). Fisher at 4%/8% *raises* classifier-judged harm above the unprotected INT3 floor (0.34/0.32 vs. 0.16; +0.18 [+0.06, +0.30] and +0.16 [+0.05, +0.27], CIs exclude 0) and above even FP16 (0.23); the heuristic agrees in direction. At 60% it fails to recover on either metric and worsens XSTest (0.47 vs. 0.50).

L Second Attack Family: PAIR

To test whether the decoupling is specific to GCG’s optimized adversarial suffixes, we evaluate the JailbreakBench PAIR transfer artifacts (Chao et al., 2025), a structurally different semantic/persuasion jailbreak, on 20 conditions with the same dual metrics (Table 15). The decoupling generalizes, while the direction of the Ministral protection effect is attack-specific (neutral under PAIR, significant under GCG). The finding that transfers across attack families is the decoupling itself; treat any single attack’s ASR as attack-specific evidence.

Condition	Llama-3.1	Qwen3	GLM-4	Ministral	DeepSeek
FP16	0.42/0.14	0.89/0.21	0.80/0.29	0.99/0.47	1.00/0.28
INT3	0.87/0.44	0.96/0.06	0.76/0.13	1.00/0.39	1.00/0.07
Fisher @ 4%	0.92/0.45	1.00/0.05	0.74/0.17	1.00/0.39	1.00/0.09
Fisher @ 8%	0.88/0.40	0.99/0.03	0.71/0.16	1.00/0.40	1.00/0.11

Table 15: **Second attack family: PAIR transfer artifacts** (semantic/persuasion jailbreak, JailbreakBench), 3-bit. Each cell is heuristic non-refusal rate / classifier-judged harmful-response rate. The decoupling generalizes: the heuristic over-states classifier-judged harm by +0.43 to +0.93 at INT3 across models, and protection leaves classifier-judged ASR at the INT3 level under PAIR as under GCG (on Llama, FP16 0.14 $\rightarrow$ INT3 0.44, protected 0.40–0.45). The Ministral net-harmful direction is GCG-specific: under PAIR the Fisher–INT3 classifier deltas are +0.00 [−0.13, +0.13] (@4%) and +0.01 [−0.12, +0.15] (@8%), CIs including 0 (paper/rebuttal/pair_panel.md).

M Activation-Level Analysis

We measure the refusal-direction separation of [Arditi et al. \(2024\)](#): per model, the refusal direction is extracted from the residual stream and we record how far 100 AdvBench harmful and 100 Alpaca harmless prompts project apart along it, taking the peak separation over layers (Table 16). The separation is essentially constant across FP16, INT3, and Fisher@8% on Llama-3.1 while the refusal rate swings by 31pp. This is direct activation-level support for Proposition 3: protection changes what the model *emits* without restoring (or needing to restore) the internal refusal representation, consistent with the multi-dimensional refusal geometry documented by [Wollschläger et al. \(2025\)](#) and [Hildebrandt et al. \(2025\)](#), under which a preserved linear direction need not preserve behavior.

Condition	Refusal-dir. separation	Refusal rate
FP16	37.5	0.68
INT3	37.0	0.37
Fisher@8%	37.0	0.51

Table 16: **The refusal representation survives while refusal behavior collapses** (Llama-3.1, 3-bit). Separation of harmful vs. harmless prompts projected on the refusal direction ([Arditi et al., 2024](#)) in the residual stream (peak over layers; 100 AdvBench vs. 100 Alpaca prompts) next to the behavioral refusal rate. The linear separability is essentially unchanged (< 2% swing) while behavior swings by 31pp: quantization’s safety failure, and protection’s recovery, are downstream of an intact refusal representation.

N Recovery Ratios for Every Cell

Table 17 applies the recovery-ratio definition of §8 to every (model, scorer, budget) cell on all five metrics. Two panel-wide readings: refusal is the axis protection most often restores (every RESTORED refusal cell on Qwen3 and Llama co-occurs with a

FLOOR or NET-HARMFUL GCG cell at small budgets), and GLM-4 and DeepSeek are N/A on the safety axes because their FP16–INT3 gaps are degenerate, the quantization-robust and unaligned categories of §9.

Model	Scorer	Budget	refusal	GCG	XSTest	MMLU	PPL
Llama-3.1	Fisher	4%	+0.59 R	−0.19 NH	−0.14 NH	+0.27 P	+0.01 F
	Fisher	8%	+0.46 R	+0.34 R	−0.25 NH	+0.44 R	+0.07 F
	Fisher	60%	+1.22 R	+0.19 P	+0.74 R	+0.41 R	+0.07 F
	Magnitude	4%	+0.36 R	+0.11 P	+0.42 R	+0.00 F	−0.01 F
	Magnitude	8%	+0.32 P	+0.06 F	−0.78 NH	+0.05 F	+0.02 F
	SNIP	4%	+0.12 P	+0.23 P	−0.08 F	+0.12 P	+0.04 F
	SNIP	8%	+0.23 P	−0.13 NH	−0.86 NH	+0.12 P	+0.02 F
	SSMP	4%	+0.21 P	+0.28 P	−0.39 NH	+0.20 P	+0.01 F
Qwen3	SSMP	8%	+0.18 P	−0.13 NH	−0.95 NH	+0.10 F	−0.01 F
	Fisher	4%	+0.52 R	−0.01 F	+0.29 P	+0.18 P	+0.13 P
	Fisher	8%	+0.73 R	+0.06 F	+0.24 P	+0.22 P	+0.18 P
	Fisher	60%	+1.16 R	+0.56 R	+0.44 R	+0.28 P	+0.40 R
	Magnitude	4%	+0.02 F	−0.04 F	+0.15 P	−0.04 F	+0.02 F
	Magnitude	8%	+0.07 F	−0.11 NH	+0.30 P	+0.07 F	+0.02 F
	SNIP	4%	+0.12 P	−0.03 F	+0.41 R	+0.11 P	+0.14 P
	SNIP	8%	+0.47 R	+0.03 F	+0.35 R	+0.17 P	+0.20 P
GLM-4	SSMP	4%	+0.22 P	−0.01 F	+0.39 R	+0.10 F	+0.14 P
	SSMP	8%	+0.43 R	+0.04 F	+0.50 R	+0.14 P	+0.20 P
	Fisher	4%	−	−	−	−0.41 NH	+0.24 P
	Fisher	8%	−	−	−	−0.09 F	+0.20 P
	Fisher	60%	−	−	−	−0.32 NH	+0.32 P
	Magnitude	4%	−	−	−	+0.23 P	+0.10 P
	Magnitude	8%	−	−	−	+0.14 P	+0.14 P
	SNIP	4%	−	−	−	−0.04 F	+0.14 P
Ministral	SNIP	8%	−	−	−	−0.36 NH	+0.14 P
	SSMP	4%	−	−	−	+0.00 F	+0.15 P
	SSMP	8%	−	−	−	+0.09 F	+0.15 P
	Fisher	4%	−0.11 NH	−0.19 NH	−0.04 F	−	+0.18 P
	Fisher	8%	+0.25 P	−0.29 NH	−0.06 F	−	+0.22 P
	Fisher	60%	+0.73 R	+0.10 F	−0.26 NH	−	+0.32 P
	Magnitude	4%	−0.04 F	−0.14 NH	+0.00 F	−	+0.10 F
	Magnitude	8%	−0.04 F	−0.10 F	−0.28 NH	−	+0.17 P
DeepSeek	SNIP	4%	+0.02 F	−0.24 NH	+0.29 P	−	+0.13 P
	SNIP	8%	+0.16 P	−0.29 NH	+0.53 R	−	+0.19 P
	SSMP	4%	+0.14 P	−0.19 NH	+0.61 R	−	+0.17 P
	SSMP	8%	+0.04 F	−0.33 NH	+0.62 R	−	+0.18 P
	Fisher	4%	−	−	−	−	+0.32 P
	Fisher	8%	−	−	−	−	+0.35 R
	Fisher	60%	−	−	−	−	+0.43 R
	Magnitude	4%	−	−	−	−	+0.03 F
	Magnitude	8%	−	−	−	−	+0.06 F
	SNIP	4%	−	−	−	−	+0.45 R
	SNIP	8%	−	−	−	−	+0.15 P
	SSMP	4%	−	−	−	−	+0.48 R
	SSMP	8%	−	−	−	−	+0.27 P

Table 17: **Recovery ratio** $\rho = (m_{\text{cond}} - m_{\text{INT3}})/(m_{\text{FP16}} - m_{\text{INT3}})$ for every (model, scorer, budget) cell, labeled by the fixed thresholds of §8: R(estored) $\rho \geq 1/3$; P(artial) $0.10 \leq \rho < 1/3$; F(loor) $|\rho| < 0.10$; NH (net-harmful) $\rho \leq -0.10$; − (N/A) when the $|\text{FP16} - \text{INT3}|$ gap is degenerate (< 0.05 ; < 1.0 for PPL). The four exhibits of Table 12 follow systematically from this one definition.

O Full-Ranking Correlation and Protect-Set Overlap

Table 18 and Figure 4 report the Spearman rank correlation between scorer scores over *all* blocks (computed from a 128-prompt calibration subset), plus the escape-clause correlation. SSMP↔SNIP is near-unity on every model; the universal negative escape correlation explains why Fisher’s top- β selection diverges everywhere even though its global ranking correlates highly with SSMP.

Figure 5 gives the full pairwise protect-set overlap for all scorer pairs and models, the block-level view behind Table 2. Two structures are visible. The SSMP↔SNIP cell is near 1 on every model, the two-factor equivalence class of Theorem 2. The Fisher and magnitude rows overlap far less with that pair and with each other, the two single-factor escapes; their mutual overlap is itself low because the escape-clause correlation is negative on every model (Table 18).

Model	Spearman ρ (all blocks)			Escape corr
	SSMP,SNIP	SSMP,Fish	SSMP,Mag	
Llama-3.1	0.998	0.947	-0.119	-0.406
Qwen3	0.999	0.834	0.294	-0.461
GLM-4	0.999	0.958	0.406	-0.334
Ministral	1.000	0.947	-0.194	-0.677
DeepSeek	0.997	0.940	-0.087	-0.544

Table 18: **Full-ranking validation of Theorem 2.** Spearman ρ between scorer scores over *all* blocks (not just top- β). SSMP $\leftrightarrow$ SNIP correlation is *near-unity* (0.997–1.000): the two-factor pair is interchangeable even globally. Globally SSMP $\leftrightarrow$ Fisher is also high (0.83–0.96) because the gradient-norm factor dominates the bulk ranking; the divergence is concentrated in the top- β protect set (Table 2). *Escape corr* = Spearman(Fisher, magnitude) on the top-8% blocks is *negative on all five models*: the highest-gradient blocks are not the highest-magnitude ones, so Fisher selects a different protect-set everywhere (its downstream *effect* depends on the model category).

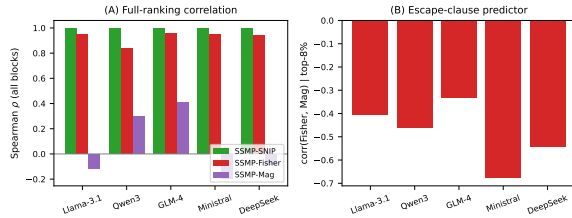


Figure 4: (A) Spearman rank correlation of each scorer with SSMP over all blocks: SSMP $\leftrightarrow$ SNIP is near-unity; single-factor scorers correlate less. (B) Escape-clause correlation $\text{corr}(\text{Fisher}, \text{magnitude})$ on the top-8% blocks: negative on all 5 models, so Fisher’s magnitude-free ranking selects different blocks everywhere.

P XSTest and Budget-Scaling Figures

Figure 6 maps every condition by its two XSTest components, compliance on safe prompts and refusal on unsafe prompts; the safety score of Table 1 is their harmonic mean. Figure 7 sweeps the Fisher protection budget from the unprotected INT3 floor to the FP16 ceiling on refusal rate and heuristic GCG ASR. The sweep separates the models with partial recovery (Llama-3.1, Qwen3) from GLM-4, which is flat because it is quantization-robust, and from Ministral and DeepSeek, which stay at the floor at every budget.

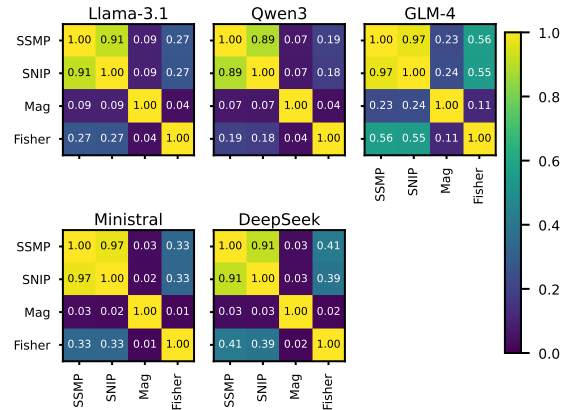


Figure 5: **Full pairwise protect-set Jaccard overlap** (top-8% blocks) for all scorer pairs, per model. SSMP $\leftrightarrow$ SNIP is near-identical (0.89–0.97); pure magnitude and Fisher each overlap far less with the two-factor pair and with each other. Random-selection baseline is 0.042. Direct block-level confirmation of Corollary 1.

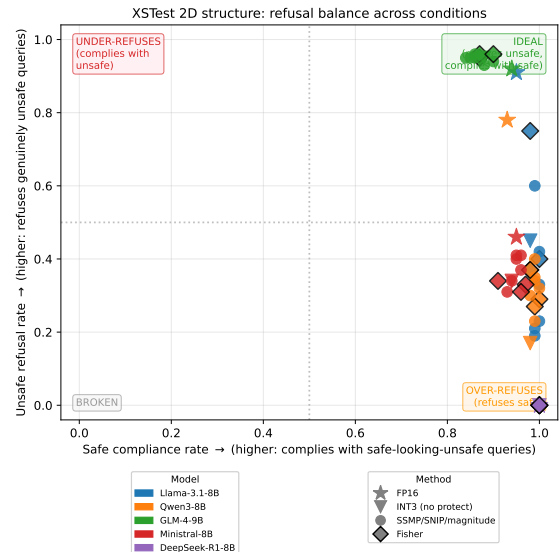


Figure 6: XSTest 2D map: safe-compliance (x) vs. unsafe-refusal (y) across all conditions. Top-right is ideal. Strongly-aligned FP16 baselines (GLM-4, Llama) sit top-right; DeepSeek conditions sit bottom-right (complies with everything); weakly-aligned protected models populate the lower-middle.

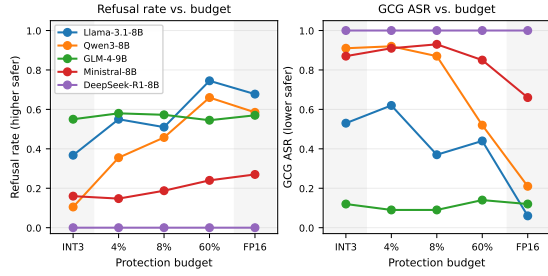


Figure 7: Refusal rate (left) and GCG ASR (right) vs. Fisher protection budget. The x -axis sweeps from the unprotected INT3 *floor* (shaded, left) through Fisher budgets 4%/8%/60% to the FP16 *ceiling* (shaded, right). Strongly-aligned models (Llama, Qwen3) show partial recovery; GLM-4 is flat (quantization-robust); Ministral and DeepSeek stay at the floor.

Q Paired-Bootstrap Statistics

Because greedy decoding is deterministic, each condition’s refusal rate is exact; statistical uncertainty arises only from the finite evaluation-prompt sample. Table 19 gives paired-over-prompts bootstrap CIs (10k resamples) for the SSMP-vs-other refusal-rate differences at $\beta=8\%$, with prompt indices resampled identically for the paired scorers. As an additional check on the scoring/selection step (which *is* seed-dependent), we recomputed the unprotected references on Qwen3 across seeds $\{0, 1, 2\}$: FP16 and INT3 refusal rates varied by $\leq 1.0\text{pp}$, confirming the selection step contributes negligible variance relative to the prompt-sampling CIs. We did not run a full multi-seed sweep across all scorers; the protect-set agreement test (Table 2, Fig. 5) provides seed-independent confirmation of the equivalence class.

Model	Comparison	$\Delta\text{refusal}$ (95% CI)
Llama-3.1	SSMP–SNIP	$-0.018 [-0.050, +0.013]$
	SSMP–Mag	$-0.045 [-0.090, +0.000]$
	SSMP–Fisher	$-0.088 [-0.128, -0.045]^*$
Qwen3	SSMP–SNIP	$-0.018 [-0.052, +0.017]$
	SSMP–Mag	$+0.172 [+0.130, +0.215]^*$
	SSMP–Fisher	$-0.145 [-0.200, -0.090]^*$
GLM-4	SSMP–SNIP	$-0.015 [-0.030, +0.000]$
	SSMP–Mag	$+0.068 [+0.042, +0.095]^*$
	SSMP–Fisher	$+0.030 [+0.007, +0.055]^*$
Ministral	SSMP–SNIP	$-0.012 [-0.025, -0.003]^*$
	SSMP–Mag	$+0.010 [-0.023, +0.042]$
	SSMP–Fisher	$-0.022 [-0.048, +0.003]$
DeepSeek	SSMP–SNIP	$+0.000 [+0.000, +0.000]$
	SSMP–Mag	$+0.000 [+0.000, +0.000]$
	SSMP–Fisher	$+0.000 [+0.000, +0.000]$

Table 19: **Paired bootstrap over evaluation prompts** (10k resamples; greedy decoding is deterministic). Refusal-rate difference of SSMP vs. each scorer at $\beta=8\%$; * marks CIs excluding 0. SSMP–SNIP is statistically indistinguishable on 4 of 5 models (Cor. 1); SSMP–Fisher is significant on the models that sit between the floor and the ceiling.

R Extended Confidence Intervals

Table 20 extends the paired-over-prompts bootstrap of Table 19 from refusal to the non-refusal metrics (heuristic GCG ASR and XSTest unsafe-refusal), computed from the stored per-prompt outcome vectors with 10k resamples; MMLU uses Wilson intervals ($n=200$), quoted where used (Exhibit A: $0.245 [0.19, 0.31]$). Per-condition intervals for every metric and cell are released with the code (paper/rebuttal/per_condition_cis.csv). The exhibit-critical deltas survive: the Qwen3 Fisher@8% refusal-vs-GCG recovery gap is $+0.68 [+0.55, +0.82]$, and the Ministral classifier-judged net-harmful deltas exclude zero (§8).

Model	Comparison	ΔGCG (95% CI)	$\Delta\text{XSTest}_{\text{unsafe}}$
Llama-3.1	SSMP–SNIP	+0.000 [−0.060, +0.060]	−0.020 [−0.080, +0.030]
	SSMP–Mag	+0.090 [+0.010, +0.170]*	−0.040 [−0.140, +0.060]
	SSMP–Fisher	+0.220 [+0.130, +0.310]*	−0.180 [−0.270, −0.090]*
Qwen3	SSMP–SNIP	−0.010 [−0.060, +0.040]	+0.080 [−0.020, +0.180]
	SSMP–Mag	−0.110 [−0.170, −0.050]*	+0.100 [−0.020, +0.220]
	SSMP–Fisher	+0.010 [−0.050, +0.070]	+0.130 [+0.010, +0.250]*
GLM-4	SSMP–SNIP	−0.020 [−0.050, +0.000]	−0.010 [−0.030, +0.000]
	SSMP–Mag	−0.030 [−0.070, +0.000]	+0.010 [−0.020, +0.050]
	SSMP–Fisher	+0.010 [−0.020, +0.040]	−0.010 [−0.030, +0.000]
Ministral	SSMP–SNIP	+0.010 [+0.000, +0.030]	+0.010 [−0.020, +0.050]
	SSMP–Mag	+0.050 [−0.010, +0.110]	+0.100 [+0.020, +0.180]*
	SSMP–Fisher	+0.010 [−0.040, +0.060]	+0.080 [+0.020, +0.140]*
DeepSeek	SSMP–SNIP	+0.000 [+0.000, +0.000]	+0.000 [+0.000, +0.000]
	SSMP–Mag	+0.000 [+0.000, +0.000]	+0.000 [+0.000, +0.000]
	SSMP–Fisher	+0.000 [+0.000, +0.000]	+0.000 [+0.000, +0.000]

Table 20: **Paired-over-prompts bootstrap CIs on the non-refusal metrics** (extends Table 19): difference of SSMP vs. each scorer at $\beta=8\%$ on heuristic GCG ASR ($n=100$) and XSTest unsafe-refusal (stratified, 100 safe + 100 unsafe); * marks CIs excluding 0. SSMP–SNIP contains 0 on every metric and model except a 1pp Ministral XSTest gap; SSMP–Fisher is significant away from the floor and the ceiling (Llama GCG +0.22 [+0.13, +0.31]), the escape pattern of Theorem 2.

S Relation to Published Retraining and High-Budget Baselines

Q-resafe (Chen et al., 2025) (SNIP scoring followed by LoRA-DPO retraining) and CWP (Al Hakim et al., 2026) (a safety–general Fisher difference at $\sim 60\%$ budget) each pair a *scoring rule* with an orthogonal *post-scoring step*: alignment retraining for Q-resafe, a large protection budget for CWP. Our study isolates the scoring rule, the component Theorem 2 characterizes, and evaluates it directly. Q-resafe’s SNIP scoring lies in the SNIP/Taylor equivalence class, which we test head-to-head (Tables 2, 18: SSMP $\leftrightarrow$ SNIP Spearman 0.997–1.000). CWP’s Fisher-family scoring is represented by our Fisher ablation, including the high-budget Fisher@60% point (the general-Fisher subtraction aside; it is a per-group constant that does not change block ranking).

Full Q-resafe pipeline run. To test whether the retraining step changes the conclusions of the scoring analysis, we additionally ran the full Q-resafe pipeline end-to-end on Llama-3.1 at 3-bit, on a single 48 GB GPU: SNIP scoring; DPO-pair construction pairing each harmful prompt’s refusal (chosen) with the quantized model’s generation (rejected); LoRA adapters attached to the SNIP-critical modules; and the DPO patch, with the frozen-reference log-probabilities precomputed once. The DPO converged (50 steps, final loss 0.012, mean chosen vs. rejected log-probability −17.0 vs. −300.1; the

published recipe uses 200 steps), and the patched model’s HarmBench refusal rate is 0.53 ($N=128$), up from unprotected INT3’s 0.37: roughly half the FP16 (0.68)–INT3 (0.37) gap, landing in the same range as the strongest small-budget scorer (Fisher@8%, 0.51). The full pipeline therefore recovers refusal comparably to scoring-only protection, consistent with the scoring analysis; its recovered refusal is measured on the same benign-refusal metric and is subject to the same decoupling caution as §8. The run configuration is released (configs/m2_qresafe_llama31.yaml). Quantifying how alignment retraining interacts with structural collapse beyond this single-model check remains a direction for future work.

T Reproducibility Statement

All scoring, protection, quantization, and evaluation code, the five model identifiers, the datasets, and the exact run configurations are released.