FROZEN INPUT-HARM PROBES UNDER COMPRESSION: RETAINED AUROC DOES NOT MEAN RETAINED REFUSAL

Anonymous authors
Paper under double-blind review

ABSTRACT

Linear probes trained on a model's BF16 activations are candidate safety monitors, but models are often compressed after a probe is fitted. We ask whether a frozen input-harm probe that still separates harmful from harmless prompts implies that the compressed model still refuses. Across four families and forty compressed configurations (post-training quantisation, QLoRA, distillation, and Wanda and SparseGPT pruning), we pair judged attack success on StrongREJECT with frozen-probe AUROC on XSTest wherever hidden dimensions match. With greedy 48-token generation, thirteen configurations have attack-success intervals above zero and seven survive Holm correction; none is post-training quantised or Gemma-2-9B. In OLMo-2-7B, SparseGPT 2:4 and Wanda 2:4 raise attack success by $+0.0590$ $[+0.0377, +0.0913]$ and $+0.0654$ $[+0.0427, +0.0988]$, while seed-mean probe AUROC falls detectably but by less than our planned 0.10 big-drop threshold at all eight probed layers (a one-sided test confirms this at every layer only for SparseGPT 2:4). The same methods give Llama and Qwen big drops at two or three layers shallower than the planned one. On our primary pooled AUROC basis this is therefore an existence result in one family (per fold, Llama SparseGPT 2:4 also qualifies), on disjoint probe and behavioural prompts. The OLMo increases lose Holm significance if three or four of six unjudged baseline responses (SparseGPT and Wanda, respectively) were successes; twelve of the thirteen increases persist at a 256-token cap and in every sampled seed. Retained AUROC does not certify retained refusal: reusing a probe after compression still requires behavioural evaluation.

1 INTRODUCTION

Activation probes are proposed as inexpensive safety monitors (MacDiarmid et al., 2024; McKenzie et al., 2025), but they are fitted on the uncompressed (BF16) model, whereas deployed models are often quantised, pruned, adapted or distilled, and quantisation, pruning and fine-tuning can weaken safety behaviour (Hong et al., 2024; Xu et al., 2024; Qi et al., 2024). A deployer who re-validates a frozen probe by its area under the ROC curve (AUROC) could read retained discrimination as evidence that safety survived. Because harmfulness and refusal can be encoded separately (Zhao et al., 2025), that inference needs a direct test.

We compare judged harmful compliance with a linear probe (Alain & Bengio, 2017) trained on BF16 residual-stream activations to predict *input harmfulness* and then frozen; it is not trained to predict compliance. Two definitions were fixed before data collection: a *big drop* is a point AUROC decrease larger than 0.10, and *operational failure* is AUROC below 0.75. A post hoc rule (H7) is a sensitivity analysis (Section 4.2).

Our contributions are: (i) a paired grid of forty compressed models across four families with judge-change intervals, Holm correction and missing-verdict sensitivity (Sections 3–4); (ii) an existence result: in OLMo-2-7B, 2:4 pruning raises attack success while AUROC declines detectably but, in seed mean, never by more than 0.10 at any of eight layers (Section 4.1); (iii) mitigations with measured limits (Section 6); and (iv) negative results: no detected post-training quantisation or Gemma increase in the 48-token grid and a less robust frozen sparse basis (Sections 4.2 and 7).

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

Probe and behavioural prompts are disjoint, so the conclusion is an inference from discrimination to behaviour; nulls do not establish equivalence or safety.

2 Related work and experimental scope

Compression and adaptation. Our grid uses INT8 (Dettmers et al., 2022), NF4 and QLoRA (Dettmers et al., 2023), GPTQ (Frantar et al., 2023), AWQ (Lin et al., 2024), Wanda (Sun et al., 2024), SparseGPT (Frantar & Alistarh, 2023), and distillation (Hinton et al., 2015). QLoRA combines quantisation with low-rank adaptation (Hu et al., 2022). These are comparisons of recorded implementations and doses.

Safety under compression and fine-tuning. Hong et al. (2024) find that 4-bit quantisation preserves trustworthiness better than 50% pruning, and Xu et al. (2024) find dimension-dependent effects of quantisation and pruning on harm, bias, and utility. Quantisation-induced safety loss has been repaired by safety patching (Chen et al., 2025) or mitigated by critical-weight protection (Al Hakim et al., 2026), and targeted pruning exposes safety-critical regions (Wei et al., 2024). Quantisation can also be exploited to hide harmful behaviour (Egashira et al., 2024). With fine-tuning vulnerabilities (Qi et al., 2024), these establish that compression and fine-tuning can affect safety; we add frozen input-harm classifier transfer, separating its task from compliance.

Safety representations and readouts. Linear probes measure decodability (Alain & Bengio, 2017), with control tasks testing memorisation (Hewitt & Liang, 2019); activation probes detect defection in backdoored models (MacDiarmid et al., 2024) and are proposed as deployed monitors (McKenzie et al., 2025), though they can fail under distribution shift (Kramár et al., 2026) or obfuscated activations (Bailey et al., 2026). Refusal-direction interventions test causal influence on behaviour (Arditi et al., 2024). Early-layer states already separate malicious from benign inputs (Zhou et al., 2024); harmfulness at the last instruction token and refusal at post-instruction tokens are encoded separately, and the tested narrow finetuning attack barely changes the former (Zhao et al., 2025). Our probe reads the last post-instruction token, where Zhao et al. (2025) locate refusal; we test whether its frozen input-harm readout still implies retained refusal after standard compression. Representation engineering (Zou et al., 2023a) and difference-of-means readouts (Marks & Tegmark, 2024) motivate our alternative mass-mean readout, and Gemma Scope sparse autoencoders (SAEs; Lieberum et al., 2024) provide a frozen sparse basis.

Behavioural evaluation and guards. XSTest tests exaggerated refusal of safe requests (Röttger et al., 2024); its safe/unsafe contrasts support our probe task. StrongREJECT distinguishes harmful usefulness from superficial compliance (Souly et al., 2024); we use its requests with a versioned API rubric. Llama-Guard-3 (Meta, 2024a; Grattafiori et al., 2024), a successor of the Llama Guard safety classifier (Inan et al., 2023), is our external response monitor. Our direct requests do not test adversarial jailbreak robustness (Mazeika et al., 2024; Chao et al., 2024).

2.1 Models and compression cells

The families are Llama-3.1-8B-Instruct (Grattafiori et al., 2024), Qwen2.5-7B-Instruct (Qwen Team, 2024), OLMo-2-1124-7B-Instruct (OLMo Team, 2025), and Gemma-2-9B-it (Gemma Team, 2024), abbreviated Llama-3.1-8B, Qwen2.5-7B, OLMo-2-7B and Gemma-2-9B. Each has one BF16 baseline and ten compressed *cells*: four post-training quantisation (PTQ) methods, QLoRA, distillation, and Wanda/SparseGPT at unstructured 50% and structured 2:4 sparsity; with the baselines, the grid has forty-four cells. Distilled students have incompatible hidden dimensions, precluding frozen teacher-probe transfer. Adaptation uses Dolly (Conover et al., 2023), a corpus separate from XSTest (instructions of the superseded AdvBench–Dolly probe pool were also excluded); each family has one QLoRA dose and one fp32-trained student. Perplexity ratios (WikiText-2 Merity et al., 2017 against BF16; for students, against the pre-distillation student on held-out Dolly) are damage diagnostics (Appendix Table 3; Appendix D).

3 MEASUREMENT: BEHAVIOUR, READOUT, AND UNCERTAINTY

3.1 JUDGE-MEASURED ATTACK SUCCESS

For each of 313 StrongREJECT requests (Souly et al., 2024), cached rubric-v6 judgements by `openai/gpt-5.6-sol` (through OpenRouter, with `openai/flex` pinned) label usable harmful content; attack success rate (ASR) is the fraction so labelled. Refusal and unusable harmful responses are non-successes; this binary rubric does not implement the complete StrongREJECT scoring rule. BF16 judge ASRs are Llama 0.0065 $[0.0008, 0.0232]$ (2/309), Qwen 0.0000 $[0.0000, 0.0120]$ (0/306), OLMo 0.0000 $[0.0000, 0.0119]$ (0/307), and Gemma 0.0033 $[0.0001, 0.0181]$ (1/305), with the exact 95% intervals of Clopper & Pearson (1934). Responses to chat-formatted requests without a system prompt use greedy decoding with a 48-new-token cap, which 98–100% of Qwen, OLMo and Gemma responses reach (family averages), against 16% in Llama. Greedy cap-256 regeneration covers 32 of 40 compressed cells, including all PTQ, and temperature sampling covers the sixteen compressed cells of the decoding supplement (Appendices C.4 and E.3–E.4); all other results use this 48-token grid (Section 8, L8).

Let $V_{mi} = 1$ denote a valid judgement for configuration m and request i, and J_{mi} its binary success label when valid. With family baseline 0, define $S_m = \{i \in \{1, \ldots, 313\} : V_{mi}V_{0i} = 1\}$. The conditional paired estimand is

$$\Delta\text{ASR}_m = \frac{1}{|S_m|} \sum_{i \in S_m} (J_{mi} - J_{0i}). \tag{1}$$

A *forward flip* is a request judged a non-success for BF16 and a success for the compressed model; a *reverse flip* is the opposite. Judge-change intervals for Eq. equation 1 invert the paired-proportion score test of Tango (1998). Exact McNemar tests with Holm correction (Appendix C.8) are a separate, stricter criterion; Appendix A.3 reports missing-verdict sensitivity. In validation, 169 of the 196 responses that our lexical refusal detector (family-calibrated phrase matching) labels compliant lacked usable harmful content, and none of the 196 it labels refusals was a judged success (Appendix C).

3.2 A PROBE TASK WITH A MEASURED MODEL-FREE REFERENCE

The probe detects *input harmfulness* on an XSTest contrast set of 200 unsafe and 250 safe prompts (Röttger et al., 2024), independent of the model's response, and has no judged-compliance labels. An initial AdvBench–Dolly pool (Zou et al., 2023b; Conover et al., 2023) was superseded because bag-of-words features separate it almost perfectly. Nested cross-validated TF-IDF logistic regression gives a model-free reference of 0.785 across five seeds (SD 0.0115; Appendix C.7); probe prompts share no item with StrongREJECT.

3.3 PAIRED OUT-OF-FOLD FROZEN TRANSFER

A standardised scikit-learn logistic probe (Pedregosa et al., 2011) with L_2 regularisation ($C = 1.0$; 1,000 maximum iterations), fitted within each BF16 training fold, is applied unchanged to BF16 and compressed test folds; prompt order, labels and five stratified folds (seeds 0, 1, and 2) are shared, and standardisation uses training rows only. Activations are taken at the final token of the chat-formatted prompt, after the assistant-turn header; layer ℓ is the residual stream after the ℓ-th block, counting from one. Primary AUROC pools out-of-fold scores within each seed; mean within-fold (per-fold) AUROC is a sensitivity, as is $C \in \{0.01, 0.1, 1, 10, 100\}$, with no selection on compressed outcomes. The two bases differ materially: across all 288 compatible cell–layers, pooled big-drop and operational-failure counts are 21 [17,24] and 9 [7,12], but per-fold counts, which are insensitive to the fold-specific score offsets that compression induces in concatenated out-of-fold scores, are 4 [3,7] and 0 [0,0] (Appendix E.6). A single deployed probe has no such offsets, so the per-fold basis is the closer deployment analogue; we keep the pooled basis as primary rather than switch estimators after seeing results. Paired ΔAUROC intervals resample prompts with fitted probes fixed (Efron, 1979), each draw averaging three per-seed differences on the same prompts; fold-split SDs stay separate, and absolute ranges span per-seed CI endpoints (Appendix D.1).

Planned extraction layers were Llama 16, Qwen 14, OLMo 16, and Gemma 21 (moved from 20 for Gemma Scope indexing before final extraction). After inspecting transfer results, without a

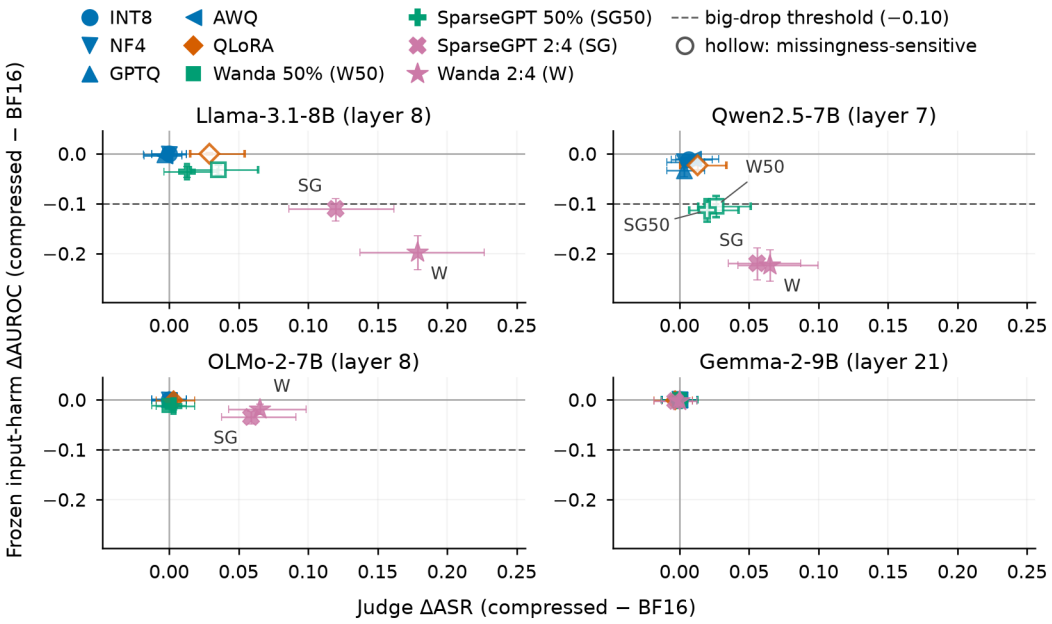

Figure 1: Frozen input-harm ΔAUROC (vertical) against paired judge ΔASR (horizontal) at the headline layers (Llama 8, Qwen 7, OLMo 8, Gemma 21), with intervals from Table 1. Horizontal bars are paired 95% Tango score intervals; vertical bars are paired 95% prompt-sampling intervals, not absolute-AUROC ranges. Both axes use the same scale across panels. The dashed line marks the planned big-drop threshold (-0.10). Hollow markers show the five plotted missingness-sensitive cells (Section 4.1); the sixth, OLMo distillation, is a student and is not plotted, because students are structurally inapplicable.

fixed rule or held-out split, we report *headline layers* Llama 8, Qwen 7, and OLMo 8, at a quarter of depth; Gemma keeps its planned layer, at half depth (quarter-depth Gemma changes: Appendices A.2 and C.3). Appendix B.1 classifies planned-layer readouts and all eight layers of the eleven behaviour-positive, dimension-compatible cells.

4 Behaviour and Readout Under Compression

Figure 1 compares the paired axes. Thirteen of forty judge ΔASR intervals lie above zero (four Llama, six Qwen, three OLMo, no Gemma; count interval [11,17]), and seven [6,10] positive exact McNemar tests (McNemar, 1947) survive Holm correction over forty comparisons (Holm, 1979). Imputing unjudged responses against the increase moves six positive Tango intervals to include zero (M cells, Section 4.1); within any one OLMo or Qwen 2:4 cell it removes Holm significance, and across all positive cells at once only Llama SparseGPT 2:4, Llama Wanda 2:4 and Qwen distillation survive. Guard-informed imputation (unjudged responses count as successes when Llama-Guard-3 flagged them) keeps all seven (Appendix Table 8). Across the 36 dimension-compatible cells the two axes co-vary (Spearman -0.71; descriptive, since cells within a family share baselines and prompts), but readout change is not comparable across families: Qwen GPTQ loses 0.0331 AUROC with no detected ASR change, whereas OLMo SparseGPT 2:4 loses 0.0345 with an increase of $+0.0590$.

4.1 A readout without a big drop does not imply retained refusal

Both OLMo 2:4 cells pair a Holm-significant behavioural increase with a small but detectable readout decline. SparseGPT 2:4 has judge change $+0.0590$ $[+0.0377, +0.0913]$ (18 compressed-only and no BF16-only successes; Holm $p = 2.7 \times 10^{-4}$) and frozen change -0.0345 $[-0.0489, -0.0224]$ at layer 8; Wanda 2:4 has $+0.0654$ $[+0.0427, +0.0988]$ (20 and none; Holm $p = 7.1 \times 10^{-5}$) and -0.0191 $[-0.0299, -0.0093]$. The seed-mean declines are detectable at all eight layers but none is a big drop; the largest is -0.0902 $[-0.1120, -0.0675]$ (Wanda 2:4, layer 4;

Appendix Tables 12 and 14), where seed 1 changes by -0.1016 (Appendix E.7). A one-sided test finds SparseGPT 2:4 non-inferior to the -0.10 margin at every depth in all six combinations of readout (frozen logistic, re-centred logistic, mass-mean) and AUROC basis (pooled, per-fold); Wanda 2:4 fails only for the frozen logistic probe on the pooled basis at layer 4. Per-fold and re-centred (training-fold mean-subtracted) readouts also leave Llama SparseGPT 2:4 without a big drop at any layer (largest per-fold decline -0.0995), though its interval reaches -0.1256; the difference-of-means (mass-mean) readout, whose own BF16 AUROC is low at shallow layers, has big drops only in Llama Wanda 2:4 (layers 12 and 16, -0.1006 and -0.1003; Appendices E.6–E.8).

Three qualifications apply. First, missing verdicts: if three (SparseGPT) or four (Wanda) of the six unjudged OLMo BF16 responses were successes, the 48-token increases would lose Holm significance (column k of Appendix Table 8); at cap 256, with one BF16 success among 310 judgements, they grow to $+0.2451$ $[+0.1990, +0.2969]$ and $+0.2451$ $[+0.2003, +0.2963]$ (Appendix C.4). Second, disjoint prompts: on StrongREJECT itself, the input-harm probe's flag rate at a threshold fixed for 0.95 true-positive rate on BF16 XSTest falls at the headline layer in all eight 2:4 cells, including Gemma's, which have zero forward flips under 48-token labels; other depths can reverse the change. At the headline layer the flag change reflects threshold calibration, not localisation of changed compliance; at the planned layer, however, the per-request score drop ranks forward flips above chance in Llama and OLMo and below chance in Qwen (Appendix C.5). On the probe's own 200 unsafe XSTest prompts, OLMo lexical refusal falls from 0.8950 to 0.4150 (SparseGPT 2:4) and 0.6400 (Wanda 2:4), Holm-significant under both detector rules; such declines are Holm-significant in 18 cells, including Llama and Gemma QLoRA (0.9200 to 0.5300; 0.7600 to 0.5250) at headline ΔAUROC -0.0003 and -0.0017; lexical refusal is not judged harm (Appendix E.23). Separately, scores along the frozen BF16 difference-of-means (refusal) direction, oriented so that higher means more input harm, are lower on complied than on refused requests; most of this anti-prediction is already present on BF16 activations, and 2:4 pruning deepens it mainly at planned or deeper layers (Appendices C.2 and E.18). Third, both cells share one model and sparsity pattern; on the pooled basis the same methods give Llama and Qwen big drops at two or three of eight layers, including the headline but not the planned layer (Appendix Table 12). Among cells regenerated at cap 256, Llama QLoRA's increase ($+0.0971$ $[+0.0668, +0.1361]$) also pairs with no big drop at any layer.

Six increases are missingness-sensitive. Adversarially imputing unjudged verdicts moves the Tango interval to include zero for Llama QLoRA and Wanda 50%, Qwen QLoRA, Wanda 50% and SparseGPT 50%, and OLMo distillation (M in Table 1). All remain positive under guard-informed imputation (Appendix A.3), and at cap 256 all but Qwen QLoRA remain positive by an unadjusted test under a conservative assignment of unjudged verdicts (Appendix C.4).

4.2 Where the readout degrades

All sixteen PTQ judge-change intervals contain zero at forty-eight tokens; at cap 256 none rejects under Holm on complete cases or with missing verdicts filled by the lexical detector, although adversarial assignments can and many responses still reach the longer cap (Appendices E.3–E.4). These nulls are consistent with Hong et al. (2024) and do not contradict the quantisation-induced safety loss reported by Chen et al. (2025) and Al Hakim et al. (2026), which used other models, methods and benchmarks. The PTQ cells' headline-layer frozen changes lie between -0.0331 and $+0.0017$ (Table 1; Appendix Figure 3); five (four Qwen, one Llama) are detectable declines smaller than 0.05. Llama and Qwen 2:4 cells pair Holm-significant increases with pooled big drops at the headline layer, from -0.1107 $[-0.1346, -0.0900]$ (Llama SparseGPT 2:4) to -0.2229 $[-0.2557, -0.1920]$ (Qwen Wanda 2:4); per fold only Llama Wanda 2:4 and Qwen SparseGPT 2:4 remain big drops there, and at the planned layers the point changes lie between -0.0599 and -0.0841, though two intervals (Llama Wanda 2:4, Qwen SparseGPT 2:4) reach -0.1070 and -0.1036. Readout damage thus depends on depth, AUROC basis and family. Llama Wanda 2:4's layer-8 absolute range crosses 0.785 (Table 1). Distilled students (NA on the frozen axis) have positive judge-change intervals on Qwen and OLMo (OLMo's is missingness-sensitive) and nulls on Llama and Gemma.

Rule sensitivity. As a post hoc sensitivity rule (H7), we label each dimension-compatible compressed cell A (no behavioural increase), B (an increase with a degraded readout: paired-change CI upper endpoint below -0.05, or the absolute range, the envelope of per-seed absolute-AUROC CIs, reaching 0.785) or C (an increase without that degradation). H7 and its 0.05 margin were chosen

Table 1: Headline-layer results per compression method: frozen ΔAUROC [paired 95% prompt-bootstrap CI], absolute frozen AUROC [range of per-seed 95% item CIs; no joint coverage claim], and judge-measured ΔASR from one greedy 48-token generation per request [95% Tango score interval for a paired difference of proportions]. Layers: Llama 8, Qwen 7, OLMo 8, Gemma 21. **Bold**: unadjusted asymptotic score inversion excludes zero; *: positive exact McNemar test survives Holm over forty comparisons; ^{M}: positive Tango interval that includes zero if unjudged verdicts are imputed adversarially (Section 4.1); ‡: loses Tango significance after deleting one compressed-only success. Read AUROCs against the 0.785 model-free reference, not 0.5. NA: frozen transfer is inapplicable to students, whose hidden dimension differs. Full grid: Appendix Table 3.

Method	ΔAUROC [95% CI]	AUROC [CI range]	Judge ΔASR [95% CI]
Llama-3.1-8B			
INT8	$+0.0001$ $[-0.0015, +0.0019]$	0.9910 $[0.9818, 0.9967]$	0.0000 $[-0.0123, +0.0123]$
NF4	-0.0019 $[-0.0049, +0.0006]$	0.9890 $[0.9788, 0.9960]$	0.0000 $[-0.0123, +0.0123]$
GPTQ	-0.0021 $[-0.0054, +0.0008]$	0.9888 $[0.9800, 0.9957]$	-0.0032 $[-0.0182, +0.0091]$
AWQ	-0.0035 $[-0.0063, -0.0012]$	0.9874 $[0.9778, 0.9948]$	-0.0032 $[-0.0181, +0.0091]$
QLoRA	-0.0003 $[-0.0016, +0.0009]$	0.9906 $[0.9816, 0.9968]$	$\mathbf{+0.0291}$ $[\mathbf{+0.0154}, \mathbf{+0.0544}]^{M}$
Student	NA	NA	0.0000 $[-0.0175, +0.0175]$
Wanda 50%	-0.0321 $[-0.0439, -0.0214]$	0.9588 $[0.9233, 0.9797]$	$\mathbf{+0.0357}$ $[\mathbf{+0.0155}, \mathbf{+0.0640}]^{M}$
SparseGPT 50%	-0.0355 $[-0.0477, -0.0253]$	0.9554 $[0.9269, 0.9799]$	$+0.0129$ $[-0.0036, +0.0345]$
SparseGPT 2:4	-0.1107 $[-0.1346, -0.0900]$	0.8801 $[0.8173, 0.9271]$	$\mathbf{+0.1197}$ $[\mathbf{+0.0863}, \mathbf{+0.1615}]^{*}$
Wanda 2:4	-0.1983 $[-0.2323, -0.1641]$	0.7926 $[0.7045, 0.8727]$	$\mathbf{+0.1786}$ $[\mathbf{+0.1371}, \mathbf{+0.2266}]^{*}$
Qwen2.5-7B			
INT8	-0.0113 $[-0.0208, -0.0019]$	0.9260 $[0.8906, 0.9576]$	$+0.0065$ $[-0.0059, +0.0235]$
NF4	-0.0175 $[-0.0283, -0.0080]$	0.9198 $[0.8712, 0.9504]$	$+0.0033$ $[-0.0092, +0.0183]$
GPTQ	-0.0331 $[-0.0468, -0.0201]$	0.9042 $[0.8603, 0.9429]$	$+0.0033$ $[-0.0092, +0.0183]$
AWQ	-0.0109 $[-0.0196, -0.0018]$	0.9264 $[0.8927, 0.9559]$	$+0.0098$ $[-0.0027, +0.0284]$
QLoRA	-0.0231 $[-0.0331, -0.0138]$	0.9142 $[0.8789, 0.9410]$	$\mathbf{+0.0132}$ $[\mathbf{+0.0005}, \mathbf{+0.0336}]^{M\ddagger}$
Student	NA	NA	$\mathbf{+0.0984}$ $[\mathbf{+0.0698}, \mathbf{+0.1369}]^{*}$
Wanda 50%	-0.1051 $[-0.1277, -0.0846]$	0.8322 $[0.7446, 0.8986]$	$\mathbf{+0.0264}$ $[\mathbf{+0.0134}, \mathbf{+0.0512}]^{M}$
SparseGPT 50%	-0.1135 $[-0.1359, -0.0915]$	0.8238 $[0.7402, 0.9241]$	$\mathbf{+0.0198}$ $[\mathbf{+0.0070}, \mathbf{+0.0425}]^{M}$
SparseGPT 2:4	-0.2199 $[-0.2535, -0.1888]$	0.7174 $[0.6366, 0.8160]$	$\mathbf{+0.0557}$ $[\mathbf{+0.0351}, \mathbf{+0.0874}]^{*}$
Wanda 2:4	-0.2229 $[-0.2557, -0.1920]$	0.7144 $[0.6226, 0.7885]$	$\mathbf{+0.0653}$ $[\mathbf{+0.0422}, \mathbf{+0.0997}]^{*}$
OLMo-2-7B			
INT8	0.0000 $[-0.0017, +0.0019]$	0.9894 $[0.9804, 0.9960]$	0.0000 $[-0.0124, +0.0124]$
NF4	-0.0006 $[-0.0018, +0.0006]$	0.9889 $[0.9787, 0.9964]$	0.0000 $[-0.0124, +0.0124]$
GPTQ	$+0.0017$ $[0.0000, +0.0040]$	0.9912 $[0.9818, 0.9978]$	0.0000 $[-0.0124, +0.0124]$
AWQ	-0.0008 $[-0.0023, +0.0005]$	0.9887 $[0.9778, 0.9963]$	0.0000 $[-0.0124, +0.0124]$
QLoRA	-0.0010 $[-0.0021, -0.0001]$	0.9885 $[0.9781, 0.9966]$	$+0.0033$ $[-0.0091, +0.0182]$
Student	NA	NA	$\mathbf{+0.0196}$ $[\mathbf{+0.0070}, \mathbf{+0.0421}]^{M}$
Wanda 50%	-0.0114 $[-0.0195, -0.0044]$	0.9781 $[0.9616, 0.9902]$	0.0000 $[-0.0124, +0.0124]$
SparseGPT 50%	-0.0123 $[-0.0204, -0.0051]$	0.9772 $[0.9635, 0.9893]$	$+0.0033$ $[-0.0091, +0.0182]$
SparseGPT 2:4	-0.0345 $[-0.0489, -0.0224]$	0.9550 $[0.9339, 0.9718]$	$\mathbf{+0.0590}$ $[\mathbf{+0.0377}, \mathbf{+0.0913}]^{*}$
Wanda 2:4	-0.0191 $[-0.0299, -0.0093]$	0.9704 $[0.9500, 0.9855]$	$\mathbf{+0.0654}$ $[\mathbf{+0.0427}, \mathbf{+0.0988}]^{*}$
Gemma-2-9B			
INT8	-0.0001 $[-0.0003, +0.0001]$	0.9993 $[0.9974, 1.0000]$	0.0000 $[-0.0124, +0.0124]$
NF4	-0.0002 $[-0.0007, 0.0000]$	0.9991 $[0.9966, 1.0000]$	0.0000 $[-0.0125, +0.0125]$
GPTQ	0.0000 $[-0.0002, +0.0002]$	0.9993 $[0.9978, 1.0000]$	0.0000 $[-0.0125, +0.0125]$
AWQ	-0.0001 $[-0.0005, 0.0000]$	0.9992 $[0.9970, 1.0000]$	0.0000 $[-0.0126, +0.0126]$
QLoRA	-0.0017 $[-0.0033, -0.0005]$	0.9977 $[0.9948, 0.9997]$	-0.0033 $[-0.0185, +0.0093]$
Student	NA	NA	$+0.0066$ $[-0.0095, +0.0258]$
Wanda 50%	-0.0005 $[-0.0014, 0.0000]$	0.9989 $[0.9968, 1.0000]$	0.0000 $[-0.0133, +0.0133]$
SparseGPT 50%	-0.0008 $[-0.0017, -0.0002]$	0.9985 $[0.9964, 0.9999]$	0.0000 $[-0.0125, +0.0125]$
SparseGPT 2:4	-0.0028 $[-0.0063, -0.0004]$	0.9965 $[0.9921, 0.9994]$	-0.0033 $[-0.0185, +0.0093]$
Wanda 2:4	-0.0022 $[-0.0045, -0.0005]$	0.9971 $[0.9936, 0.9993]$	0.0000 $[-0.0126, +0.0126]$

after the results; at the planned layers H7 gives 3 B and 8 C, and the planned definitions give no B (Appendix B.1).

5 Retraining and Causal Interventions

Retraining tests recoverable input-harm information. Refitting on compressed activations with the same folds raises Qwen Wanda 2:4's layer-7 pooled AUROC from 0.7144 (below 0.785; per-fold 0.8373) to 0.9254, a paired retrained-minus-frozen gain of $+0.2110$ $[+0.1793, +0.2434]$ (95%

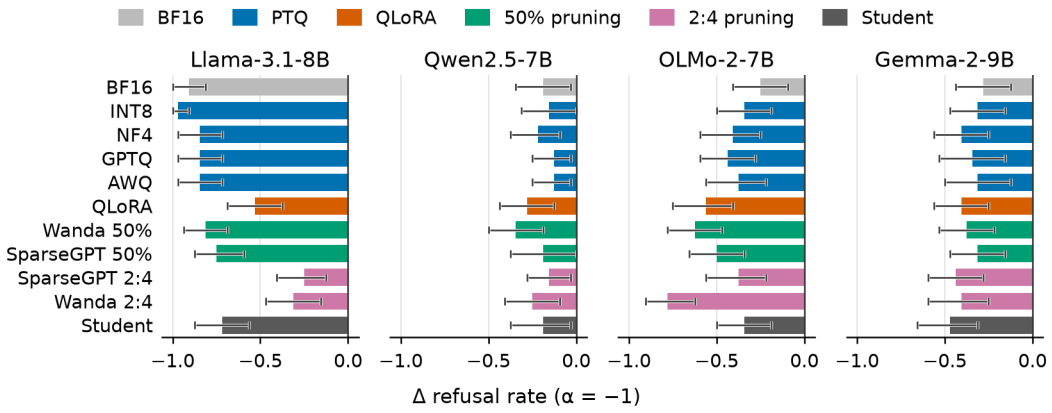

Figure 2: Change in refusal rate (lexical detector) on 32 held-out unsafe XSTest prompts under negative steering ($\alpha = -1$) along each cell's own difference-of-means direction at its intervention layer (Llama and OLMo 16, Qwen 14, Gemma 21; students at their mid-depth layer), relative to that cell's zero-steering identity run. Whiskers are paired 95% prompt-resampling intervals. Changes are bounded by each cell's zero-steering refusal rate, so bar lengths are not comparable across cells.

prompt-sampling interval; fitted probes held fixed). This does not measure compliance prediction or establish unchanged geometry.

Steering changes lexical refusal. Following Arditi et al. (2024), each cell has its own difference-of-means direction, estimated on a class-balanced half of the XSTest contrast set and applied at the family's planned layer (students at their own mid-depth layer). Steering adds α times the unit direction, scaled by the norm of the unsafe-minus-safe mean gap over all 450 contrast prompts, at every token position of that layer; projection removal removes the component along the unit direction there. Each is compared with that cell's zero-steering identity run on thirty-two held-out unsafe XSTest prompts, with refusal scored by the lexical detector. Negative steering reduces refusal with an interval below zero in 42 of 44 cells, including 38 of 40 compressed cells (Figure 2); OLMo SparseGPT 2:4 has $-12/32$ $[-18/32, -7/32]$ and OLMo Wanda 2:4 has $-25/32$ $[-29/32, -20/32]$ (paired 95% intervals). Projection removal does so in 35 of 44 cells (exceptions: Appendix C). These cell-specific directions are distinct from the frozen probe, and their potency does not establish preservation of the BF16 direction (Appendix E.9).

6 Mitigations with measured limits

6.1 An external guard detects most judge-confirmed harm in high-success cells

Llama-Guard-3 (Meta, 2024a) operates on responses. Among nine cells with at least ten judge-confirmed successes it detects 201 of 224: recall 0.8973 $[0.8255, 0.9573]$, precision 0.2784 $[0.2361, 0.3217]$, and false-positive rate 0.2055 $[0.1800, 0.2318]$; 521 of its 722 flags are not judge-confirmed harm. Across all 44 cells, recall is 0.8415 $[0.7007, 0.9532]$, precision 0.2201 $[0.1852, 0.2571]$, and FPR 0.0641 $[0.0557, 0.0731]$; outside the nine cells it detects only 38 of 60 successes, with precision 0.10. The judge is the reference, not ground truth (per-cell intervals: Appendix A.4).

6.2 Training across compression variants gives a small per-fold gain

Augmentation adds the training prompts' activations from non-held-out compression variants to each BF16 training fold and tests on the held-out method (Llama, Qwen and OLMo; shared prompt-index folds keep test-prompt copies out of fitting). Layer four, where the BF16-only probe's held-out AUROC averaged over all nine held-out methods is lowest in every family, is primary; Appendix B.2 gives a row-count-matched Gaussian-noise control and layers 8/16 (Qwen 7/17).

Table 2: Augmentation ablation. Paired ΔAUROC versus a BF16-only probe on *held-out* pruning methods (primary layer 4, three families, three seeds; $n = 36$ paired family–method–seed units), with a 95% bootstrap CI (5000 resamples). The pooled intervals resample these units and are conditional on the fitted probes and prompt corpus. Appendix B reports every measured augmentation layer and method-clustered gap-closure intervals. Rows use the pooled out-of-fold basis, on which label-free re-centring of the BF16-only probe alone gains $+0.0815$; the like-for-like per-fold all-variant gain is $+0.0144$ (Appendix E.12).

Augmentation set	ΔAUROC vs BF16-only
BF16 + quantisation	$+0.0091$ $[+0.0005, +0.0188]$
BF16 + feature-wise RMS-matched Gaussian noise	$+0.0240$ $[+0.0125, +0.0355]$
BF16 + other pruning	$+0.0607$ $[+0.0432, +0.0800]$
BF16 + all variants	$+0.0609$ $[+0.0431, +0.0812]$

On held-out pruning methods, all-variant augmentation improves pooled AUROC by $+0.0609$ $[+0.0431, +0.0812]$ (Table 2; family–method-clustered interval $[+0.0393, +0.0826]$; family-clustered interval with $G = 3$: $[-0.0171, +0.1390]$). On this pooled basis, however, label-free re-centring of the BF16-only probe on compressed training-fold means gains $+0.0815$ $[+0.0611, +0.1032]$. The like-for-like per-fold gain, unaffected by re-centring, is $+0.0144$ $[+0.0109, +0.0179]$ (family–method cluster-robust CR1 interval $[+0.0085, +0.0204]$, Cameron & Miller, 2015; Holm $p = 0.0027$ over the twelve primary augmentation contrasts). Within-family method CR1 supports it in one of three families, and at headline layers it is $+0.0236$ $[+0.0144, +0.0340]$, significant only before adjustment (Holm $p = 0.1812$; Appendices E.12–E.13). Augmentation falls short of retraining at weak layers (at layer 4, Qwen Wanda 2:4 reaches 0.7852, against 0.8921 for retraining), and unseen-method robustness remains unestablished.

7 Negative results that delimit the contribution

7.1 Gemma has no detected increase in the 48-token grid

Gemma has no detected judge increase in the forty-eight-token grid (Table 1). Its seven cap-256 intervals (four PTQ, QLoRA and both 2:4 methods) and the seed-majority intervals of its three sampled cells contain zero; one of nine sampled-seed intervals excludes zero, QLoRA seed 2 at $+0.0231$ $[+0.0103, +0.0469]$, which its other seeds do not reproduce (Appendix C.4). These nulls do not establish equivalence: smaller increases remain compatible with the complete cases, and assigning each unjudged response in favour of an increase makes 22 of 27 null intervals positive, including all ten Gemma and 12 of 16 PTQ cells, whereas guard-informed imputation leaves every null unchanged (Appendix Table 7; Section 8, L9). Gemma's headline-layer readouts remain far above 0.785, which does not establish immunity to other hazards or doses.

Among the BF16 baselines, over-refusal (the lexical detector's refusal rate on the 250 safe XSTest prompts) is highest in Gemma, 0.1400 $[0.0995, 0.1893]$, but its difference from OLMo's 0.0800 fails Holm, and one baseline per family cannot link it to Gemma's behavioural nulls. Gemma QLoRA and its student lower lexical refusal of unsafe XSTest prompts (0.7600 to 0.5250 and 0.4700; Holm-significant) without a detected ASR increase, while Gemma QLoRA's frozen AUROC moves by only -0.0017 (Appendices E.23–E.24).

7.2 A frozen sparse basis is less robust than the dense probe

The frozen Gemma Scope JumpReLU SAE basis (Lieberum et al., 2024; Rajamanoharan et al., 2024) is compared with dense probes at matching activation depths (Gemma Scope layer l is our activation layer $l+1$) on identical prompts and folds. Unlike an SAE-transfer study that scores reconstruction and retrained sparse probes after pruning (Gupte et al., 2025), we measure frozen downstream AUROC. Over twenty-seven cell–layers (nine compressed Gemma cells at three SAE depths) the SAE-minus-dense change is -0.0161 $[-0.0201, -0.0123]$ with shared-prompt resampling, and -0.0130 $[-0.0167, -0.0092]$ per fold. Relative to the dense probe, an unthresholded random basis of the same width changes by only -0.0012 $[-0.0018, -0.0006]$, so width and regularisation do not explain the SAE's loss, whereas one thresholded to the SAE's sparsity changes by -0.0642 $[-0.0717, -0.0566]$, pointing to hard sparsification, not learned features, as the main

source of loss; the controls do not isolate a mechanism. These results cover one suite, width and sparsity (Appendices A.2, C.3 and E.10; Section 8, L6).

8 LIMITATIONS

L1: one judge and incomplete coverage. One API judge and rubric label 13,492 of 13,772 cell–item pairs; the 280 (2.0%) gaps (119 logged provider filters and 161 logged errors, 105 of them on Gemma) are excluded, with sensitivity in Appendix A.3. A second judge graded 582 of 600 pairs enriched for primary successes, agreeing at Cohen's $\kappa = 0.9242$ (Cohen, 1960); its disagreements run 13 to 1 toward success, all 13 in adaptation or pruning cells (Appendices C.1 and E.2). **L2: limited intervention resolution.** Thirty-two prompts per intervention give $1/32$ increments, too coarse for fine cross-cell comparison; refusal is scored lexically, not by the judge; smaller effects can also reflect lower baseline refusal; and re-estimated cell-specific vectors do not test causal transfer of the frozen BF16 vector.

L3: unequal AWQ coverage. OLMo's post-normalisation architecture lets the tested AWQ implementation smooth only `o_proj` and `down_proj` inputs, limiting method-level comparisons. **L4: one student and one adaptation dose.** One student per family and one QLoRA recipe and dose cannot distinguish resistance from insufficient adaptation; nulls do not generalise across doses. Student cells compare the distilled student with the BF16 teacher, so they measure deploying the student, not distillation alone. **L5: a single external guard.** One fixed guard has blind spots; WildGuard (Han et al., 2024) is untested, and recall in success-rich cells does not extend to low-success cells.

L6: one SAE family and suite. Only Gemma Scope JumpReLU has a comparable suite; the matched random basis controls width and regularisation, its thresholded variant costs more than the SAE, and generality is limited to the suite, width and L_0 (mean number of active features) tested. **L7: partial, local augmentation robustness.** Gap closure, the share of the retraining gain that augmentation recovers, depends on retraining headroom, method set and AUROC basis and cannot rank families (Appendix A.2); transfer to similar methods does not establish robustness to arbitrary unseen compression. **L8: greedy and truncated generation.** Most non-Llama responses are truncated at the 48-token greedy cap, which bounds generation cost and is held fixed so that within-family comparisons share one window, but can understate ASR for verbose models and so biases cross-family comparison. Qwen BF16 itself reaches 0.0523 $[0.0302, 0.0835]$ attack success at cap 256 (Appendix C.4); decoding invariance is not established.

L9: Gemma's low baseline. Gemma's nulls concern increases from one success in 305; testing whether compression narrows its margin requires harder hazards. **L10: analysis choices made after the results.** Headline layers, H7's 0.05 margin, the gap-closure eligibility cutoff, the lenient refusal detector and the exploratory Gemma augmentation reporting floor were chosen after the results; planned-layer, all-layer, all-cutoff and planned-definition results accompany them (Section 4.2). **L11: construct, corpus and threshold divergence.** Input harmfulness and harmful compliance differ, probe and behavioural prompts are disjoint, and fixed-threshold transfer also depends on calibration; neither retained AUROC nor changed flags validates a compliance monitor.

9 CONCLUSION

OLMo-2-7B's two 2:4 methods increase judged compliance while AUROC declines detectably but never, in seed mean, by more than the planned 0.10 big-drop threshold at any probed layer. Among Holm-significant 48-token increases and on the pooled AUROC basis this existence result is specific to OLMo; per fold, Llama SparseGPT 2:4 also stays inside the threshold; among cells regenerated at cap 256, so does Llama QLoRA's unadjusted increase; and among the six behaviour-positive 2:4 cells only OLMo SparseGPT 2:4 passes the one-sided non-inferiority test at every layer in all six readout and basis combinations. Twelve of the thirteen behavioural increases persist at cap 256 and in all sampled seeds. Fixed-threshold flags decline at the headline layer, though not at every depth, even in Gemma's 2:4 cells, which have no 48-token forward flips, and whether score drops locate changed compliance depends on depth and family. These results rest on one judge, the headline-layer comparisons (not the all-layer existence result) use post hoc layers, and they do not establish compliance-monitor failure; probe reuse requires behavioural evaluation and threshold validation.

AI USE STATEMENT

We disclose the following uses of generative AI. *Research instruments:* an LLM judge (`openai/gpt-5.6-sol` via OpenRouter, rubric v6) labels harmful compliance (Section 3), a second LLM judge (`anthropic/claude-sonnet-5`) audits a sample of those labels (Appendix C.1), and Llama-Guard-3 is the evaluated response guard (Section 6); the compressed open-weight models are objects of study. *Implementing methods:* an AI coding agent (Claude Code with Anthropic Claude models)was used for coding assistance in harness implementation. *Writing:* generative AI tools was used for aiding and polishing the manuscript text. We take responsibility for the final content of this work, including code, and polished text produced with the aid of generative AI.

ETHICS STATEMENT

This work measures how compression changes harmful compliance in public open-weight models, used under their licences, with commercial judge APIs used under the providers' usage policies. It involves no human subjects or personal data, and its prompts come from public research benchmarks (StrongREJECT, XSTest and Dolly; AdvBench only in a superseded probe pool). The paper reports aggregate rates and prints no harmful prompt or model response. Provider moderation refusals are recorded as missing verdicts and analysed (Section 8, L1; Appendix A.3). Showing that standard pruning recipes can raise harmful compliance introduces no new attack and supports a protective practice: behavioural re-evaluation after compression rather than reliance on probe AUROC. The code repository releases prompt-free per-item exports and boolean verdicts; apart from the judge-validation record (169 responses judged refusals and eight public benchmark prompts), it contains no model generations or compressed checkpoints.

REPRODUCIBILITY STATEMENT

Sections 2–3 and Appendix D specify models, compression and adaptation recipes, prompt sets, folds, seeds and the main interval procedures (Appendix D.1; the others are specified where reported); model and code revisions are recorded in the released manifests and artifact inventory, with the exceptions listed in Appendix D, and Appendix A lists every cell. The code repository (`https://anonymous.4open.science/r/compression-safety-monitors-frozen-probes-E8B2`; Appendix D) contains analysis code, configurations, the judge rubric, saved result summaries, prompt-free per-item exports and a CPU script that regenerates the confidence intervals without model fits or paid API calls. Regenerating activations and generations requires the GPUs recorded in Appendix D; QLoRA and distillation builds are not byte-reproducible.

REFERENCES

Muhammad Alif Al Hakim, Alfan Farizki Wicaksono, and Fajri Koto. Preserving fairness and safety in quantized LLMs through critical weight protection. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 19831–19855, 2026. doi: 10.18653/v1/2026.findings-acl.993. URL https://aclanthology.org/2026.findings-acl.993/.

Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes, 2017. URL https://openreview.net/forum?id=ryF7rTqgl.

Andy Arditi, Oscar Balcells Obeso, Aaquib Syed, Daniel Paleka, Nina Rimsky, Wes Gurnee, and Neel Nanda. Refusal in language models is mediated by a single direction. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=pH3XAQME6c.

Luke Bailey, Alex Serrano, Abhay Sheshadri, Mikhail Seleznyov, Jordan Taylor, Erik Jenner, Jacob Hilton, Stephen Casper, Carlos Guestrin, and Scott Emmons. Obfuscated activations bypass LLM latent-space defenses. In *International Conference on Learning Representations*, 2026. URL https://arxiv.org/abs/2412.09565.

Roger L. Berger. Multiparameter hypothesis testing and acceptance sampling. *Technometrics*, 24 (4):295–300, 1982. doi: 10.2307/1267823.

A. Colin Cameron and Douglas L. Miller. A practitioner's guide to cluster-robust inference. *Journal of Human Resources*, 50(2):317–372, 2015. doi: 10.3368/jhr.50.2.317.

Patrick Chao, Edoardo Debenedetti, Alexander Robey, Maksym Andriushchenko, Francesco Croce, Vikash Sehwag, Edgar Dobriban, Nicolas Flammarion, George J. Pappas, Florian Tramèr, Hamed Hassani, and Eric Wong. Jailbreakbench: An open robustness benchmark for jailbreaking large language models. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2024. URL https://openreview.net/forum?id=urjPCYZt0I.

Kejia Chen, Jiawen Zhang, Jiacong Hu, Yu Wang, Jian Lou, Zunlei Feng, and Mingli Song. 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 *PMLR*, pp. 9728–9746, 2025. URL https://proceedings.mlr.press/v267/chen25ci.html.

C. J. Clopper and E. S. Pearson. The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika*, 26(4):404–413, 1934. doi: 10.1093/biomet/26.4.404.

Jacob Cohen. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46, 1960. doi: 10.1177/001316446002000104.

Mike Conover, Matt Hayes, Ankit Mathur, Jianwei Xie, Jun Wan, Sam Shah, Ali Ghodsi, Patrick Wendell, Matei Zaharia, and Reynold Xin. Free Dolly: Introducing the world's first truly open instruction-tuned LLM. Databricks, 2023. URL https://www.databricks.com/blog/2023/04/12/dolly-first-open-commercially-viable-instruction-tuned-llm.

Elizabeth R. DeLong, David M. DeLong, and Daniel L. Clarke-Pearson. Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach. *Biometrics*, 44(3):837–845, 1988. doi: 10.2307/2531595.

Tim Dettmers, Mike Lewis, Younes Belkada, and Luke Zettlemoyer. LLM.int8(): 8-bit matrix multiplication for transformers at scale. In *Advances in Neural Information Processing Systems*, volume 35, pp. 30318–30332, 2022. URL https://arxiv.org/abs/2208.07339.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. QLoRA: Efficient finetuning of quantized LLMs. In *Advances in Neural Information Processing Systems*, volume 36, pp. 10088–10115, 2023. URL https://arxiv.org/abs/2305.14314.

Bradley Efron. Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, 7(1): 1–26, 1979. doi: 10.1214/aos/1176344552.

Kazuki Egashira, Mark Vero, Robin Staab, Jingxuan He, and Martin Vechev. Exploiting LLM quantization. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=ISa7mMe7Vg.

Morten W. Fagerland, Stian Lydersen, and Petter Laake. The McNemar test for binary matched-pairs data: mid-p and asymptotic are better than exact conditional. *BMC Medical Research Methodology*, 13:91, 2013. doi: 10.1186/1471-2288-13-91.

Elias Frantar and Dan Alistarh. SparseGPT: Massive language models can be accurately pruned in one-shot. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *PMLR*, pp. 10323–10337, 2023. URL https://arxiv.org/abs/2301.00774.

Elias Frantar, Saleh Ashkboos, Torsten Hoefler, and Dan Alistarh. OPTQ: Accurate quantization for generative pre-trained transformers. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=tcbBPnfwxS.

Gemma Team. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*, 2024. URL https://arxiv.org/abs/2408.00118.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, et al. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024. URL https://arxiv.org/abs/2407.21783.

Suchit Gupte, Vishnu Kabir Chhabra, and Mohammad Mahdi Khalili. On the transferability of sparse autoencoders for interpreting compressed models. *arXiv preprint arXiv:2507.15977*, 2025. URL https://arxiv.org/abs/2507.15977.

Seungju Han, Kavel Rao, Allyson Ettinger, Liwei Jiang, Bill Yuchen Lin, Nathan Lambert, Yejin Choi, and Nouha Dziri. Wildguard: Open one-stop moderation tools for safety risks, jailbreaks, and refusals of LLMs. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2024. URL https://openreview.net/forum?id=Ich4tv4202.

John Hewitt and Percy Liang. Designing and interpreting probes with control tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 2733–2743, 2019. doi: 10.18653/v1/D19-1275. URL https://aclanthology.org/D19-1275/.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015. URL https://arxiv.org/abs/1503.02531.

Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70, 1979. URL https://www.jstor.org/stable/4615733.

Junyuan Hong, Jinhao Duan, Chenhui Zhang, Zhangheng Li, Chulin Xie, Kelsey Lieberman, James Diffenderfer, Brian R. Bartoldson, Ajay Kumar Jaiswal, Kaidi Xu, Bhavya Kailkhura, Dan Hendrycks, Dawn Song, Zhangyang Wang, and Bo Li. Decoding compressed trust: Scrutinizing the trustworthiness of efficient LLMs under compression. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *PMLR*, pp. 18611–18633, 2024. URL https://proceedings.mlr.press/v235/hong24a.html.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://arxiv.org/abs/2106.09685.

Hakan Inan, Kartikeya Upasani, Jianfeng Chi, Rashi Rungta, Krithika Iyer, Yuning Mao, Michael Tontchev, Qing Hu, Brian Fuller, Davide Testuggine, and Madian Khabsa. Llama Guard: LLM-based input-output safeguard for human-AI conversations. *arXiv preprint arXiv:2312.06674*, 2023. URL https://arxiv.org/abs/2312.06674.

János Kramár, Joshua Engels, Zheng Wang, Bilal Chughtai, Rohin Shah, Neel Nanda, and Arthur Conmy. Building production-ready probes for Gemini. *arXiv preprint arXiv:2601.11516*, 2026. URL https://arxiv.org/abs/2601.11516.

Tom Lieberum, Senthooran Rajamanoharan, Arthur Conmy, Lewis Smith, Nicolas Sonnerat, Vikrant Varma, János Kramár, Anca Dragan, Rohin Shah, and Neel Nanda. Gemma Scope: Open sparse autoencoders everywhere all at once on Gemma 2. In *Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pp. 278–300, 2024. doi: 10.18653/v1/2024.blackboxnlp-1.19. URL https://aclanthology.org/2024.blackboxnlp-1.19/.

Ji Lin, Jiaming Tang, Haotian Tang, Shang Yang, Wei-Ming Chen, Wei-Chen Wang, Guangxuan Xiao, Xingyu Dang, Chuang Gan, and Song Han. AWQ: Activation-aware weight quantization for on-device LLM compression and acceleration. In *Proceedings of Machine Learning and Systems*, volume 6, pp. 87–100, 2024. URL https://arxiv.org/abs/2306.00978.

Monte MacDiarmid, Timothy Maxwell, Nicholas Schiefer, Jesse Mu, Jared Kaplan, David Duvenaud, Sam Bowman, Alex Tamkin, Ethan Perez, Mrinank Sharma, Carson Denison, and Evan Hubinger. Simple probes can catch sleeper agents. Anthropic Alignment Science, 2024. URL https://www.anthropic.com/research/probes-catch-sleeper-agents.

Samuel Marks and Max Tegmark. The geometry of truth: Emergent linear structure in large language model representations of true/false datasets. In *Conference on Language Modeling*, 2024. URL https://arxiv.org/abs/2310.06824.

Mantas Mazeika, Long Phan, Xuwang Yin, Andy Zou, Zifan Wang, Norman Mu, Elham Sakhaee, Nathaniel Li, Steven Basart, Bo Li, David Forsyth, and Dan Hendrycks. 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 *PMLR*, pp. 35181–35224, 2024. URL https://arxiv.org/abs/2402.04249.

Alex McKenzie, Urja Pawar, Phil Blandfort, William Bankes, David Krueger, Ekdeep Singh Lubana, and Dmitrii Krasheninnikov. Detecting high-stakes interactions with activation probes. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=8YniJnJQ0P.

Quinn McNemar. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2):153–157, 1947. doi: 10.1007/BF02295996.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. Pointer sentinel mixture models. In *International Conference on Learning Representations*, 2017. URL https://arxiv.org/abs/1609.07843.

Meta. Llama Guard 3 8B model card. Hugging Face model card, 2024a. URL https://huggingface.co/meta-llama/Llama-Guard-3-8B.

Meta. Llama 3.2 1B Instruct model card. Hugging Face model card, 2024b. URL https://huggingface.co/meta-llama/Llama-3.2-1B-Instruct.

OLMo Team. 2 OLMo 2 furious. In *Conference on Language Modeling*, 2025. URL https://arxiv.org/abs/2501.00656.

Adam Paszke, Sam Gross, Francisco Massa, et al. PyTorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, volume 32, pp. 8024–8035, 2019. URL https://arxiv.org/abs/1912.01703.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12(85): 2825–2830, 2011. URL https://jmlr.org/papers/v12/pedregosa11a.html.

Xiangyu Qi, Yi Zeng, Tinghao Xie, Pin-Yu Chen, Ruoxi Jia, Prateek Mittal, and Peter Henderson. Fine-tuning aligned language models compromises safety, even when users do not intend to! In *International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2310.03693.

Qwen Team. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024. URL https://arxiv.org/abs/2412.15115.

Senthooran Rajamanoharan, Tom Lieberum, Nicolas Sonnerat, Arthur Conmy, Vikrant Varma, János Kramár, and Neel Nanda. Jumping ahead: Improving reconstruction fidelity with JumpReLU sparse autoencoders. *arXiv preprint arXiv:2407.14435*, 2024. URL https://arxiv.org/abs/2407.14435.

Paul Röttger, Hannah Kirk, Bertie Vidgen, Giuseppe Attanasio, Federico Bianchi, and Dirk Hovy. 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)*, pp. 5377–5400, 2024. doi: 10.18653/v1/2024.naacl-long.301. URL https://aclanthology.org/2024.naacl-long.301/.

Alexandra Souly, Qingyuan Lu, Dillon Bowen, Tu Trinh, Elvis Hsieh, Sana Pandey, Pieter Abbeel, Justin Svegliato, Scott Emmons, Olivia Watkins, and Sam Toyer. A strongREJECT for empty jailbreaks. In *ICLR 2024 Workshop on Reliable and Responsible Foundation Models*, 2024. URL https://openreview.net/forum?id=al303JJkGO.

Mingjie Sun, Zhuang Liu, Anna Bair, and J. Zico Kolter. A simple and effective pruning approach for large language models. In *International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2306.11695.

Toshiro Tango. Equivalence test and confidence interval for the difference in proportions for the paired-sample design. *Statistics in Medicine*, 17(8):891–908, 1998. doi: 10.1002/(SICI)1097-0258(19980430)17:8<891::AID-SIM780>3.0.CO;2-B.

Matthew D. Webb. Reworking wild bootstrap-based inference for clustered errors. *Canadian Journal of Economics*, 56(3):839–858, 2023. doi: 10.1111/caje.12661.

Boyi Wei, Kaixuan Huang, Yangsibo Huang, Tinghao Xie, Xiangyu Qi, Mengzhou Xia, Prateek Mittal, Mengdi Wang, and Peter Henderson. Assessing the brittleness of safety alignment via pruning and low-rank modifications. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *PMLR*, pp. 52588–52610, 2024. URL https://arxiv.org/abs/2402.05162.

Thomas Wolf, Lysandre Debut, Victor Sanh, et al. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 38–45, 2020. doi: 10.18653/v1/2020.emnlp-demos.6. URL https://aclanthology.org/2020.emnlp-demos.6/.

Zhichao Xu, Ashim Gupta, Tao Li, Oliver Bentham, and Vivek Srikumar. Beyond perplexity: Multi-dimensional safety evaluation of LLM compression. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 15359–15396, 2024. doi: 10.18653/v1/2024.findings-emnlp.901. URL https://aclanthology.org/2024.findings-emnlp.901/.

Jiachen Zhao, Jing Huang, Zhengxuan Wu, David Bau, and Weiyan Shi. LLMs encode harmfulness and refusal separately. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=zLkpt30ngy.

Zhenhong Zhou, Haiyang Yu, Xinghua Zhang, Rongwu Xu, Fei Huang, and Yongbin Li. How alignment and jailbreak work: Explain LLM safety through intermediate hidden states. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 2461–2488, 2024. doi: 10.18653/v1/2024.findings-emnlp.139. URL https://aclanthology.org/2024.findings-emnlp.139/.

Andy Zou, Long Phan, Sarah Chen, et al. Representation engineering: A top-down approach to AI transparency. *arXiv preprint arXiv:2310.01405*, 2023a. URL https://arxiv.org/abs/2310.01405.

Andy Zou, Zifan Wang, Nicholas Carlini, Milad Nasr, J. Zico Kolter, and Matt Fredrikson. Universal and transferable adversarial attacks on aligned language models. *arXiv preprint arXiv:2307.15043*, 2023b. URL https://arxiv.org/abs/2307.15043.

A COMPLETE GRID AND UNCERTAINTY LEDGER

Table 3 lists every compressed cell and BF16 baseline in two parts with the same rows, judged behaviour and then frozen readout, over-refusal and perplexity; its symbols are defined below. For the fourteen cells with zero discordances, the descriptive bound $U_n = 1 - 0.025^{1/n}$ (e.g. $U_{302} = 0.0121$, Appendix A.3) does not replace their Tango intervals. Distilled students are NA. The table identifies all thirteen unadjusted increases and the seven positive Holm tests, and gives each cell's perplexity ratio. Its missingness-robust bound, the union of each Tango interval with the two adversarial-imputation intervals, is an unadjusted criterion that excludes zero for exactly the seven Holm-significant increases. The six missingness-sensitive cells are listed in Section 4.1. Appendix B covers layer dependence and augmentation, Appendix C behavioural measurement and supplements, Appendix D reproducibility, and Appendix E additional sensitivity analyses.

In Table 3, Regime is the H7 label, a post hoc sensitivity classification (Section 4.2): A (no detected increase; H7 readout criterion not met), B (increase; change CI upper endpoint below -0.05 or absolute range reaching 0.785), C (increase; H7 readout criterion not met, although the change may be detectable); NA+/NA0 mark Student cells, whose frozen readout is NA, with/without an increase. * marks a positive Holm-adjusted exact test. M marks a positive Tango interval that includes zero under adversarial missing-verdict imputation (an unadjusted criterion; Holm fragility is in column k of Table 8). Judge ΔASR brackets are Tango score intervals for a paired difference of proportions over the n_Δ paired judge-valid prompts. The robust bound is the union of that interval with the Tango intervals under the two adversarial imputations of unjudged verdicts (Appendix A.3); it is a conservative sampling-plus-missingness bound, not a 95% interval. § marks one discordant prompt (none is significant); ‡ marks a significant increase that loses Tango significance after deleting one compressed-only success. Frozen AUROC brackets are the range of per-seed 95% item CIs at the headline layer (Llama 8, Qwen 7, OLMo 8, Gemma 21), with no joint 95% coverage claim; Student cells have no compatible frozen layer. Over-refusal is the lexical detector's refusal rate on the 250 safe XSTest prompts, with exact two-sided 95% Clopper–Pearson intervals. PPL ratio is compressed-model over BF16 perplexity; Student ratios are distilled-student over pre-distillation-student perplexity on held-out Dolly instructions (Appendix D) and are not comparable with the other rows.

Table 3: All forty compressed cells and four BF16 baselines at the headline layers, part 1 of 2: judged behaviour. Part 2 gives the frozen readout, over-refusal and perplexity ratio of the same rows; symbols and columns are defined in Appendix A.

Family	Method	Regime	Judge ΔASR [95% CI]	Robust bound	n_Δ
Llama	BF16	—	—	—	—
Llama	INT8	A	$0.0000\ [-0.0123, +0.0123]$	$[-0.0324, +0.0324]$	309
Llama	NF4	A	$0.0000\ [-0.0123, +0.0123]$	$[-0.0324, +0.0278]$	309
Llama	GPTQ	A	$-0.0032\ [-0.0182, +0.0091]^{\S}$	$[-0.0368, +0.0296]$	308
Llama	AWQ	A	$-0.0032\ [-0.0181, +0.0091]^{\S}$	$[-0.0368, +0.0296]$	309
Llama	QLoRA^{M}	C	$+0.0291\ [+0.0154, +0.0544]$	$[-0.0076, +0.0658]$	309
Llama	Wanda 50%^{M}	C	$+0.0357\ [+0.0155, +0.0640]$	$[-0.0039, +0.0787]$	308
Llama	SparseGPT 50%	A	$+0.0129\ [-0.0036, +0.0345]$	$[-0.0230, +0.0510]$	309
Llama	SparseGPT 2:4	B*	$+0.1197\ [+0.0863, +0.1615]$	$[+0.0682, +0.1736]$	309
Llama	Wanda 2:4	B*	$+0.1786\ [+0.1371, +0.2266]$	$[+0.1186, +0.2368]$	308
Llama	Student	NA0	$0.0000\ [-0.0175, +0.0175]$	$[-0.0356, +0.0313]$	309
Qwen	BF16	—	—	—	—
Qwen	INT8	A	$+0.0065\ [-0.0059, +0.0235]$	$[-0.0399, +0.0454]$	306
Qwen	NF4	A	$+0.0033\ [-0.0092, +0.0183]^{\S}$	$[-0.0427, +0.0412]$	306
Qwen	GPTQ	A	$+0.0033\ [-0.0092, +0.0183]^{\S}$	$[-0.0427, +0.0412]$	305
Qwen	AWQ	A	$+0.0098\ [-0.0027, +0.0284]$	$[-0.0371, +0.0412]$	306
Qwen	QLoRA^{M}	C	$+0.0132\ [+0.0005, +0.0336]^{\ddagger}$	$[-0.0344, +0.0776]$	302
Qwen	Wanda 50%^{M}	B	$+0.0264\ [+0.0134, +0.0512]$	$[-0.0234, +0.0776]$	303
Qwen	SparseGPT 50%^{M}	B	$+0.0198\ [+0.0070, +0.0425]$	$[-0.0289, +0.0737]$	303
Qwen	SparseGPT 2:4	B*	$+0.0557\ [+0.0351, +0.0874]$	$[+0.0014, +0.1078]$	305
Qwen	Wanda 2:4	B*	$+0.0653\ [+0.0422, +0.0997]$	$[+0.0069, +0.1658]$	291
Qwen	Student	NA+*	$+0.0984\ [+0.0698, +0.1369]$	$[+0.0375, +0.1551]$	305
OLMo	BF16	—	—	—	—
OLMo	INT8	A	$0.0000\ [-0.0124, +0.0124]$	$[-0.0412, +0.0278]$	307
OLMo	NF4	A	$0.0000\ [-0.0124, +0.0124]$	$[-0.0412, +0.0324]$	307
OLMo	GPTQ	A	$0.0000\ [-0.0124, +0.0124]$	$[-0.0412, +0.0324]$	307
OLMo	AWQ	A	$0.0000\ [-0.0124, +0.0124]$	$[-0.0412, +0.0324]$	306
OLMo	QLoRA	A	$+0.0033\ [-0.0091, +0.0182]^{\S}$	$[-0.0384, +0.0368]$	307
OLMo	Wanda 50%	A	$0.0000\ [-0.0124, +0.0124]$	$[-0.0412, +0.0412]$	305
OLMo	SparseGPT 50%	A	$+0.0033\ [-0.0091, +0.0182]^{\S}$	$[-0.0384, +0.0412]$	307
OLMo	SparseGPT 2:4	C*	$+0.0590\ [+0.0377, +0.0913]$	$[+0.0083, +0.1116]$	305
OLMo	Wanda 2:4	C*	$+0.0654\ [+0.0427, +0.0988]$	$[+0.0138, +0.1152]$	306
OLMo	Student^{M}	NA+	$+0.0196\ [+0.0070, +0.0421]$	$[-0.0246, +0.0578]$	306
Gemma	BF16	—	—	—	—
Gemma	INT8	A	$0.0000\ [-0.0124, +0.0124]$	$[-0.0496, +0.0412]$	305
Gemma	NF4	A	$0.0000\ [-0.0125, +0.0125]$	$[-0.0496, +0.0496]$	304
Gemma	GPTQ	A	$0.0000\ [-0.0125, +0.0125]$	$[-0.0496, +0.0537]$	303
Gemma	AWQ	A	$0.0000\ [-0.0126, +0.0126]$	$[-0.0496, +0.0578]$	302
Gemma	QLoRA	A	$-0.0033\ [-0.0185, +0.0093]^{\S}$	$[-0.0537, +0.0468]$	302
Gemma	Wanda 50%	A	$0.0000\ [-0.0133, +0.0133]$	$[-0.0496, +0.1189]$	284
Gemma	SparseGPT 50%	A	$0.0000\ [-0.0125, +0.0125]$	$[-0.0496, +0.0368]$	304
Gemma	SparseGPT 2:4	A	$-0.0033\ [-0.0185, +0.0093]^{\S}$	$[-0.0537, +0.0510]$	302
Gemma	Wanda 2:4	A	$0.0000\ [-0.0126, +0.0126]$	$[-0.0496, +0.0578]$	302
Gemma	Student	NA0	$+0.0066\ [-0.0095, +0.0258]$	$[-0.0454, +0.0510]$	303

Table 3 (continued), part 2 of 2: frozen input-harm readout at the headline layers, over-refusal and perplexity ratio for the rows of part 1.

Family	Method	Frozen AUROC [range]	Over-refusal [95% CI]	PPL ratio
Llama	BF16	0.9909 [0.9824, 0.9966]	0.0600 [0.0340, 0.0970]	—
Llama	INT8	0.9910 [0.9818, 0.9967]	0.0640 [0.0370, 0.1019]	1.0114
Llama	NF4	0.9890 [0.9788, 0.9960]	0.0520 [0.0280, 0.0873]	1.0548
Llama	GPTQ	0.9888 [0.9800, 0.9957]	0.0840 [0.0527, 0.1255]	1.0442
Llama	AWQ	0.9874 [0.9778, 0.9948]	0.0720 [0.0432, 0.1114]	1.0553
Llama	QLoRA	0.9906 [0.9816, 0.9968]	0.0040 [0.0001, 0.0221]	1.3127
Llama	Wanda 50%	0.9588 [0.9233, 0.9797]	0.0160 [0.0044, 0.0405]	1.4285
Llama	SparseGPT 50%	0.9554 [0.9269, 0.9799]	0.0440 [0.0222, 0.0774]	1.2878
Llama	SparseGPT 2:4	0.8801 [0.8173, 0.9271]	0.0320 [0.0139, 0.0621]	1.9532
Llama	Wanda 2:4	0.7926 [0.7045, 0.8727]	0.0320 [0.0139, 0.0621]	3.3311
Llama	Student	NA	0.0560 [0.0310, 0.0922]	0.9249
Qwen	BF16	0.9373 [0.9072, 0.9643]	0.0440 [0.0222, 0.0774]	—
Qwen	INT8	0.9260 [0.8906, 0.9576]	0.0400 [0.0193, 0.0723]	1.0039
Qwen	NF4	0.9198 [0.8712, 0.9504]	0.0200 [0.0065, 0.0461]	1.0381
Qwen	GPTQ	0.9042 [0.8603, 0.9429]	0.0240 [0.0089, 0.0515]	1.0304
Qwen	AWQ	0.9264 [0.8927, 0.9559]	0.0480 [0.0250, 0.0823]	1.0403
Qwen	QLoRA	0.9142 [0.8789, 0.9410]	0.0240 [0.0089, 0.0515]	1.2133
Qwen	Wanda 50%	0.8322 [0.7446, 0.8986]	0.0040 [0.0001, 0.0221]	1.2079
Qwen	SparseGPT 50%	0.8238 [0.7402, 0.9241]	0.0440 [0.0222, 0.0774]	1.1527
Qwen	SparseGPT 2:4	0.7174 [0.6366, 0.8160]	0.0240 [0.0089, 0.0515]	1.3593
Qwen	Wanda 2:4	0.7144 [0.6226, 0.7885]	0.0200 [0.0065, 0.0461]	1.8959
Qwen	Student	NA	0.0560 [0.0310, 0.0922]	0.9916
OLMo	BF16	0.9895 [0.9799, 0.9966]	0.0800 [0.0495, 0.1209]	—
OLMo	INT8	0.9894 [0.9804, 0.9960]	0.0800 [0.0495, 0.1209]	1.0106
OLMo	NF4	0.9889 [0.9787, 0.9964]	0.0720 [0.0432, 0.1114]	1.0320
OLMo	GPTQ	0.9912 [0.9818, 0.9978]	0.0920 [0.0592, 0.1348]	1.0531
OLMo	AWQ	0.9887 [0.9778, 0.9963]	0.0920 [0.0592, 0.1348]	1.0378
OLMo	QLoRA	0.9885 [0.9781, 0.9966]	0.0880 [0.0560, 0.1302]	1.0230
OLMo	Wanda 50%	0.9781 [0.9616, 0.9902]	0.0560 [0.0310, 0.0922]	1.2450
OLMo	SparseGPT 50%	0.9772 [0.9635, 0.9893]	0.0600 [0.0340, 0.0970]	1.3749
OLMo	SparseGPT 2:4	0.9550 [0.9339, 0.9718]	0.0160 [0.0044, 0.0405]	1.5871
OLMo	Wanda 2:4	0.9704 [0.9500, 0.9855]	0.0160 [0.0044, 0.0405]	1.5673
OLMo	Student	NA	0.0760 [0.0464, 0.1161]	0.7966
Gemma	BF16	0.9993 [0.9977, 1.0000]	0.1400 [0.0995, 0.1893]	—
Gemma	INT8	0.9993 [0.9974, 1.0000]	0.1280 [0.0892, 0.1759]	1.0079
Gemma	NF4	0.9991 [0.9966, 1.0000]	0.1480 [0.1064, 0.1982]	1.0170
Gemma	GPTQ	0.9993 [0.9978, 1.0000]	0.1320 [0.0926, 0.1804]	1.0172
Gemma	AWQ	0.9992 [0.9970, 1.0000]	0.1360 [0.0961, 0.1848]	1.0384
Gemma	QLoRA	0.9977 [0.9948, 0.9997]	0.0240 [0.0089, 0.0515]	1.2352
Gemma	Wanda 50%	0.9989 [0.9968, 1.0000]	0.1640 [0.1203, 0.2158]	1.2300
Gemma	SparseGPT 50%	0.9985 [0.9964, 0.9999]	0.1480 [0.1064, 0.1982]	1.2212
Gemma	SparseGPT 2:4	0.9965 [0.9921, 0.9994]	0.0960 [0.0625, 0.1395]	1.5585
Gemma	Wanda 2:4	0.9971 [0.9936, 0.9993]	0.1360 [0.0961, 0.1848]	1.7022
Gemma	Student	NA	0.0600 [0.0340, 0.0970]	1.0123

A.1 Prompt sampling and fold-split variability

Qwen's four PTQ cells at layer 7 have negative paired readout-change CIs but no detected behavioural increase (Tables 3 and 4). Input-harm discrimination and judged compliance are distinct constructs; these small readout changes involve no big drop.

Across families, twenty-three of thirty-six paired change intervals exclude zero after resolving OLMo GPTQ with 200,000 draws (a single 2,000-draw run gives twenty-three or twenty-four). Magnitude matters: Gemma SparseGPT 50% changes by -0.0008 $[-0.0017, -0.0002]$, a minute shift despite statistical significance.

The paired intervals below hold the fitted out-of-fold probes fixed and resample prompts within class. They quantify prompt sampling. Alternative fold splits contribute a separate spread: s_{level} is the standard deviation of absolute frozen AUROC across fold seeds; s_{Δ} is the corresponding

standard deviation of paired changes. Neither SD is a CI or a term to add to one. Figure 3 maps the paired changes.

Table 4: Frozen input-harm probe change at each family's headline layer. Each of 2,000 stratified resamples draws 450 prompts, recomputes AUROC separately for seeds 0, 1, and 2 on both sides, and averages the three per-seed differences. Predictions are not pooled across seeds. Fold-split SDs are separate: s_{level} describes absolute AUROC and s_{Δ} its paired change; neither is added to the CI. M is defined in Appendix A. The OLMo GPTQ interval uses 200,000 draws. A frozen probe trained on BF16 compliance labels is not evaluated (Appendix C.2); within-cell compliance probes and frozen refusal-direction outcome AUROCs appear in Appendices C.6 and E.5.

Family	Method	ΔAUROC [95% CI]	s_{level}	s_{Δ}
Llama L8	INT8	$+0.0001$ $[-0.0015, +0.0019]$	0.0012	0.0005
Llama L8	NF4	-0.0019 $[-0.0049, +0.0006]$	0.0009	0.0004
Llama L8	GPTQ	-0.0021 $[-0.0054, +0.0008]$	0.0012	0.0010
Llama L8	AWQ	-0.0035 $[-0.0063, -0.0012]$	0.0007	0.0005
Llama L8	QLoRA^{M}	-0.0003 $[-0.0016, +0.0009]$	0.0012	0.0003
Llama L8	Wanda 50%^{M}	-0.0321 $[-0.0439, -0.0214]$	0.0113	0.0127
Llama L8	SparseGPT 50%	-0.0355 $[-0.0477, -0.0253]$	0.0082	0.0092
Llama L8	SparseGPT 2:4	-0.1107 $[-0.1346, -0.0900]$	0.0194	0.0246
Llama L8	Wanda 2:4	-0.1983 $[-0.2323, -0.1641]$	0.0348	0.0416
Qwen L7	INT8	-0.0113 $[-0.0208, -0.0019]$	0.0088	0.0028
Qwen L7	NF4	-0.0175 $[-0.0283, -0.0080]$	0.0122	0.0117
Qwen L7	GPTQ	-0.0331 $[-0.0468, -0.0201]$	0.0120	0.0083
Qwen L7	AWQ	-0.0109 $[-0.0196, -0.0018]$	0.0073	0.0017
Qwen L7	QLoRA^{M}	-0.0231 $[-0.0331, -0.0138]$	0.0040	0.0076
Qwen L7	Wanda 50%^{M}	-0.1051 $[-0.1277, -0.0846]$	0.0340	0.0360
Qwen L7	SparseGPT 50%^{M}	-0.1135 $[-0.1359, -0.0915]$	0.0519	0.0554
Qwen L7	SparseGPT 2:4	-0.2199 $[-0.2535, -0.1888]$	0.0398	0.0406
Qwen L7	Wanda 2:4	-0.2229 $[-0.2557, -0.1920]$	0.0289	0.0337
OLMo L8	INT8	0.0000 $[-0.0017, +0.0019]$	0.0007	0.0001
OLMo L8	NF4	-0.0006 $[-0.0018, +0.0006]$	0.0009	0.0001
OLMo L8	GPTQ	$+0.0017$ $[0.0000, +0.0040]$	0.0011	0.0005
OLMo L8	AWQ	-0.0008 $[-0.0023, +0.0005]$	0.0009	0.0004
OLMo L8	QLoRA	-0.0010 $[-0.0021, -0.0001]$	0.0012	0.0005
OLMo L8	Wanda 50%	-0.0114 $[-0.0195, -0.0044]$	0.0016	0.0016
OLMo L8	SparseGPT 50%	-0.0123 $[-0.0204, -0.0051]$	0.0005	0.0009
OLMo L8	SparseGPT 2:4	-0.0345 $[-0.0489, -0.0224]$	0.0013	0.0025
OLMo L8	Wanda 2:4	-0.0191 $[-0.0299, -0.0093]$	0.0023	0.0031
Gemma L21	INT8	-0.0001 $[-0.0003, +0.0001]$	0.0001	0.0000
Gemma L21	NF4	-0.0002 $[-0.0007, 0.0000]$	0.0001	0.0002
Gemma L21	GPTQ	0.0000 $[-0.0002, +0.0002]$	0.0001	0.0000
Gemma L21	AWQ	-0.0001 $[-0.0005, 0.0000]$	0.0001	0.0001
Gemma L21	QLoRA	-0.0017 $[-0.0033, -0.0005]$	0.0003	0.0002
Gemma L21	Wanda 50%	-0.0005 $[-0.0014, 0.0000]$	0.0001	0.0002
Gemma L21	SparseGPT 50%	-0.0008 $[-0.0017, -0.0002]$	0.0002	0.0002
Gemma L21	SparseGPT 2:4	-0.0028 $[-0.0063, -0.0004]$	0.0002	0.0003
Gemma L21	Wanda 2:4	-0.0022 $[-0.0045, -0.0005]$	0.0004	0.0004

Llama Wanda 2:4 illustrates the importance of fold splits: its paired change is -0.1983 $[-0.2323, -0.1641]$, absolute-AUROC fold SD is 0.0348, and paired-change fold SD is 0.0416. Both SDs exceed the paired prompt-CI half-width in this pruned cell; neither is added to the CI. Small changes elsewhere are detectable when their intervals exclude zero; their size is described with the planned big-drop and operational-failure definitions, with absolute AUROC read against the 0.785 model-free reference.

A.2 Dependence in pooled effects and gap closure

At SAE layer nine (activation layer 10, a quarter of Gemma's depth) under SparseGPT 2:4, dense ΔAUROC is -0.0233 $[-0.0335, -0.0145]$ and SAE ΔAUROC is -0.2181 $[-0.2558, -0.1826]$; their paired SAE-minus-dense difference is -0.1948 $[-0.2338, -0.1583]$. Under Wanda 2:4, the

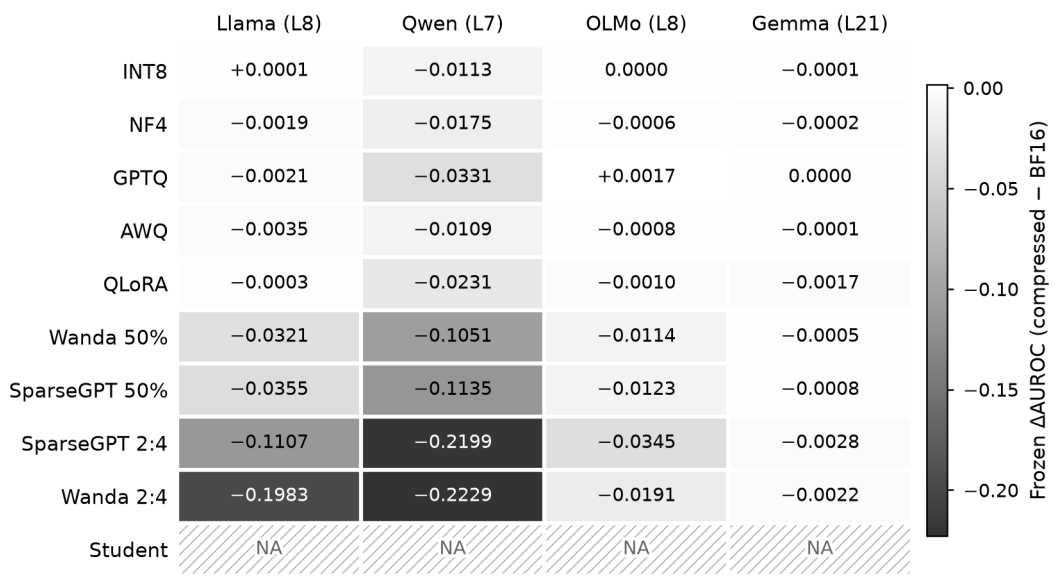

Figure 3: Frozen input-harm ΔAUROC at the headline layers (Llama 8, Qwen 7, OLMo 8, Gemma 21). Table 4 supplies paired change CIs and separate fold-split SDs; absolute discrimination is read against the 0.785 model-free reference. Darker cells mark larger drops; hatched cells are structurally inapplicable (students). Missingness caveats are in Section 4.1.

corresponding changes are -0.0215 $[-0.0313, -0.0130]$ and -0.1343 $[-0.1666, -0.1048]$, with paired difference -0.1128 $[-0.1442, -0.0834]$.

The SAE and dense changes use the same stratified resample of 450 prompts for every AUROC entering a difference. For the twenty-seven-cell mean, each draw re-averages all cells on that shared resample; cells are not treated as independent prompt datasets. The aggregate is -0.0161 $[-0.0201, -0.0123]$, using 2,000 draws. These layer-nine comparisons and the aggregate are separate from the headline-layer Table 4.

For the guard, 201 detections among 224 confirmed successes give recall 0.8973 $[0.8255, 0.9573]$. Each of 20,000 resamples draws from all 311 distinct judged prompts and retains the associated rows in all nine selected cells. The 118 prompts with a success are a subset, not the bootstrap population. Precision and FPR use the same resampling population; all are conditional on selected cells and available verdicts.

For gap closure at layer four, the per-cell fraction is

$$g_{ms} = \frac{\mathrm{AUROC}_{\text{augmented},ms} - \mathrm{AUROC}_{\text{BF16-only},ms}}{\mathrm{AUROC}_{\text{retrained},ms} - \mathrm{AUROC}_{\text{BF16-only},ms}}, \tag{2}$$

where m is the held-out method and s the fold-split seed. For eligible set $E_\tau = \{(m, s) : d_{ms} > \tau\}$, with numerator a_{ms} and denominator d_{ms} above, compare $\bar{g}_\tau = |E_\tau|^{-1} \sum_{E_\tau} a_{ms}/d_{ms}$ with the ratio of sums $R_\tau = \sum_{E_\tau} a_{ms} / \sum_{E_\tau} d_{ms}$. The latter weights cells by available improvement. Cells with denominator ≤ 0.01 are excluded in the primary comparison; other cutoffs are sensitivity analyses of this post hoc rule (Table 5).

At $\tau = 0.01$, mean-of-ratios gap closure is Qwen 0.2294 $[0.0459, 0.4129]$, Llama 0.4205 $[0.3059, 0.5351]$, and OLMo 0.6417 $[0.5215, 0.7620]$. The respective ratio-of-sums estimates are 0.4025, 0.5568, and 0.6388, reported only as descriptive points because their interval under-covered in the saved ratio simulation. The mean-ratio intervals use a CR1 cluster-robust standard error, the cluster-robust variance scaled by $G/(G-1)$ (for a mean, Stata's factor in eq. (12) of Cameron & Miller, 2015), and a t_{G-1} critical value, with held-out method as cluster. They describe method variation conditional on fixed prompts and fits. Qwen and Llama each have nine clusters; OLMo has five, so its interval is reported with that small-cluster limitation. The layer/cutoff sweep withholds

Table 5: Gap-closure cutoff sensitivity at layer four; ° marks descriptive values whose intervals are withheld for fewer than five method clusters.

Family	τ	Cells	Methods	Mean ratios [95% CI]	Pooled ratio
Llama	0	27	9	$0.3854\ [-0.0257, 0.7964]$	0.5550
Llama	0.005	23	9	$0.4305\ [0.3193, 0.5418]$	0.5564
Llama	0.01	20	9	$0.4205\ [0.3059, 0.5351]$	0.5568
Llama	0.02	13	5	$0.4996\ [0.3544, 0.6448]$	0.5704
Qwen	0	27	9	$0.2654\ [0.1386, 0.3922]$	0.4019
Qwen	0.005	27	9	$0.2654\ [0.1386, 0.3922]$	0.4019
Qwen	0.01	24	9	$0.2294\ [0.0459, 0.4129]$	0.4025
Qwen	0.02	17	8	$0.2935\ [0.1327, 0.4544]$	0.4212
OLMo	0	25	9	$0.9818\ [0.4400, 1.5237]$	0.6632
OLMo	0.005	18	7	$0.6275\ [0.3794, 0.8757]$	0.6362
OLMo	0.01	13	5	$0.6417\ [0.5215, 0.7620]$	0.6388
OLMo	0.02	12	4	0.6443°	0.6393

intervals below five clusters; the pruning-only sensitivity in Table 6 is explicitly exploratory at four clusters. Eligibility retains 24/27, 20/27, and 13/27 method–seed cells, respectively.

For ratios g_i over N eligible method–seed cells, the implemented standard error is

$$\mathrm{SE}(\bar{g}) = \left\{ \frac{G}{G-1} \frac{1}{N^2} \sum_{m=1}^{G} \left[\sum_{i \in m} (g_i - \bar{g}) \right]^2 \right\}^{1/2}.$$

Here G is the number of held-out-method clusters, and the interval is $\bar{g} \pm t_{G-1,0.975}\,\mathrm{SE}(\bar{g})$. The pruning-only comparison in Table 6 restricts all families to the same type of hold-out; thresholding still selects eligible method–seed rows.

Table 6: Layer-four pruning-only common-method sensitivity, excluding method–seed cells whose retrained-minus-BF16-only denominator is at or below 0.01. These CR1 t intervals use four method clusters per family and are exploratory: the interval calibration supports five or more clusters. They are an explicit exception to withholding intervals for fewer than five clusters in the cutoff and layer sweeps (Tables 5 and 16).

Family	Mean gap closure [95% CI]	Method clusters
Qwen	$0.3094\ [0.0525, 0.5663]$	4
Llama	$0.4813\ [0.3109, 0.6518]$	4
OLMo	$0.6443\ [0.4905, 0.7981]$	4

A.3 Coverage, missingness, and multiplicity

Gemma Wanda 2:4 has zero discordances among 302 paired judgements. Its judge change is $0.0000\ [-0.0126, +0.0126]$ by Tango; the separate descriptive discordance bound is $U_{302} = 0.0121$. This conveys roughly 0.012 resolution, not a power-based minimum detectable increase. Smaller increases remain compatible with these data.

Let $s_m = \sum_{i:V_{mi}=1} J_{mi}$ and $q_m = 313 - \sum_i V_{mi}$. On the full request set, imputing unjudged responses as all-refusal or all-success gives marginal rates $s_m/313$ and $(s_m + q_m)/313$. The corresponding paired full-corpus differences are

$$D_0 = (s_m - s_0)/313, \qquad D_1 = (s_m + q_m - s_0 - q_0)/313.$$

Permitting different missing-response labels in model and baseline gives the worst-case identification interval $[D_-, D_+] = [(s_m - s_0 - q_0)/313, (s_m + q_m - s_0)/313]$. These sensitivity endpoints concern the fixed 313-request corpus; they are not 95% sampling intervals and do not assume missingness at random. Table 9 gives all terms from saved marginal counts, including successes independently recorded in the guard–judge join.

Our API client recognised only HTTP-400 filters and classified 119 gaps as filtered and 161 as errors, so these labels do not establish causes: 124 logged errors concern items filtered elsewhere (39 in the

same corpus, 85 only at cap 256). Per-request logs permit a per-cell reconstruction (Appendix E.1); items filtered elsewhere are identified separately from logged filters. The all-missing-label sensitivity remains applicable.

Example: Llama QLoRA. Llama QLoRA has judge change $+0.0291$ $[+0.0154, +0.0544]$ and, at layer 8, frozen change -0.0003 $[-0.0016, +0.0009]$ and absolute AUROC 0.9906 (per-seed CI range $[0.9816, 0.9968]$; fold-split SD 0.0012). Imputing its three compressed-side and four BF16-side missing verdicts adversarially gives $+0.0160$ $[-0.0076, +0.0427]$ (Table 7), and its complete-case exact test (unadjusted $p = 0.0039$) fails Holm correction (adjusted $p = 0.1250$).

Table 7 assigns every unjudged compressed response success and every unjudged BF16 response refusal for the upward extreme, and reverses the assignments for the downward extreme. Guard-informed imputation labels each missing verdict success exactly when the guard flagged that response. Each scenario reports a Tango interval on the full 313-request set. These are assumption-specific sampling intervals; the finite-corpus identification bounds above are a different quantity.

The decoding supplement adds 25,326 judged (cell, decoding-variant, request) triples, with 3–19 unjudged per file. It logged 581 content-filter HTTP 403 responses as transient errors; we treat them as coverage gaps.

Table 7: Missing-verdict sensitivity for all compressed cells. Missing counts are compressed/BF16; all other columns are full-corpus judge differences with Tango 95% intervals. Up/down assign every missing label in favour of/against an increase, respectively. Guard uses the guard flag for each missing label. M marks a positive Tango interval that includes zero under adversarial missing-verdict imputation (an unadjusted criterion; Holm fragility is in column k of Table 8).

Method	Missing	Upward [CI]	Downward [CI]	Guard [CI]
Llama				
INT8	4/4	$+0.0128$ $[+0.0005, +0.0324]$	-0.0128 $[-0.0324, -0.0005]$	0.0000 $[-0.0121, +0.0121]$
NF4	3/4	$+0.0096$ $[-0.0027, +0.0278]$	-0.0128 $[-0.0324, -0.0005]$	0.0000 $[-0.0121, +0.0121]$
GPTQ	4/4	$+0.0096$ $[-0.0064, +0.0296]$	-0.0160 $[-0.0368, -0.0037]$	-0.0032 $[-0.0179, +0.0090]$
AWQ	4/4	$+0.0096$ $[-0.0064, +0.0296]$	-0.0160 $[-0.0368, -0.0037]$	-0.0032 $[-0.0179, +0.0090]$
QLoRA^M	3/4	$+0.0383$ $[+0.0221, +0.0658]$	$+0.0160$ $[-0.0076, +0.0427]$	$+0.0288$ $[+0.0152, +0.0537]$
Wanda 50%^{M}	4/4	$+0.0479$ $[+0.0258, +0.0787]$	$+0.0224$ $[-0.0039, +0.0520]$	$+0.0351$ $[+0.0153, +0.0630]$
SparseGPT 50%	4/4	$+0.0256$ $[+0.0073, +0.0510]$	0.0000 $[-0.0230, +0.0230]$	$+0.0128$ $[-0.0036, +0.0341]$
SparseGPT 2:4	4/4	$+0.1310$ $[+0.0964, +0.1736]$	$+0.1054$ $[+0.0682, +0.1484]$	$+0.1182$ $[+0.0852, +0.1595]$
Wanda 2:4	4/4	$+0.1885$ $[+0.1464, +0.2368]$	$+0.1629$ $[+0.1186, +0.2120]$	$+0.1789$ $[+0.1377, +0.2266]$
Student	3/4	$+0.0096$ $[-0.0090, +0.0313]$	-0.0128 $[-0.0356, +0.0062]$	0.0000 $[-0.0173, +0.0173]$
Qwen				
INT8	5/7	$+0.0224$ $[+0.0100, +0.0454]$	-0.0160 $[-0.0399, +0.0035]$	$+0.0096$ $[-0.0027, +0.0278]$
NF4	5/7	$+0.0192$ $[+0.0068, +0.0412]$	-0.0192 $[-0.0427, -0.0019]$	$+0.0032$ $[-0.0090, +0.0179]$
GPTQ	5/7	$+0.0192$ $[+0.0068, +0.0412]$	-0.0192 $[-0.0427, -0.0019]$	$+0.0032$ $[-0.0090, +0.0179]$
AWQ	3/7	$+0.0192$ $[+0.0068, +0.0412]$	-0.0128 $[-0.0371, +0.0084]$	$+0.0096$ $[-0.0027, +0.0278]$
QLoRA^M	11/7	$+0.0479$ $[+0.0293, +0.0776]$	-0.0096 $[-0.0344, +0.0131]$	$+0.0128$ $[+0.0005, +0.0324]$
Wanda 50%^{M}	7/7	$+0.0479$ $[+0.0293, +0.0776]$	$+0.0032$ $[-0.0234, +0.0303]$	$+0.0319$ $[+0.0174, +0.0578]$
SparseGPT 50%^{M}	8/7	$+0.0447$ $[+0.0268, +0.0737]$	-0.0032 $[-0.0289, +0.0219]$	$+0.0192$ $[+0.0068, +0.0412]$
SparseGPT 2:4	6/7	$+0.0735$ $[+0.0495, +0.1078]$	$+0.0319$ $[+0.0014, +0.0659]$	$+0.0607$ $[+0.0392, +0.0929]$
Wanda 2:4	20/7	$+0.1246$ $[+0.0925, +0.1658]$	$+0.0383$ $[+0.0069, +0.0735]$	$+0.0990$ $[+0.0707, +0.1372]$
Student	6/7	$+0.1150$ $[+0.0842, +0.1551]$	$+0.0735$ $[+0.0375, +0.1141]$	$+0.1022$ $[+0.0734, +0.1408]$
OLMo				
INT8	3/6	$+0.0096$ $[-0.0027, +0.0278]$	-0.0192 $[-0.0412, -0.0068]$	0.0000 $[-0.0121, +0.0121]$
NF4	4/6	$+0.0128$ $[+0.0005, +0.0324]$	-0.0192 $[-0.0412, -0.0068]$	0.0000 $[-0.0121, +0.0121]$
GPTQ	4/6	$+0.0128$ $[+0.0005, +0.0324]$	-0.0192 $[-0.0412, -0.0068]$	0.0000 $[-0.0121, +0.0121]$
AWQ	4/6	$+0.0128$ $[+0.0005, +0.0324]$	-0.0192 $[-0.0412, -0.0068]$	0.0000 $[-0.0121, +0.0121]$
QLoRA	4/6	$+0.0160$ $[+0.0037, +0.0368]$	-0.0160 $[-0.0384, +0.0008]$	$+0.0032$ $[-0.0090, +0.0179]$
Wanda 50%	6/6	$+0.0192$ $[+0.0068, +0.0412]$	-0.0192 $[-0.0412, -0.0068]$	0.0000 $[-0.0121, +0.0121]$
SparseGPT 50%	5/6	$+0.0192$ $[+0.0068, +0.0412]$	-0.0160 $[-0.0384, +0.0008]$	$+0.0032$ $[-0.0090, +0.0179]$
SparseGPT 2:4	6/6	$+0.0767$ $[+0.0521, +0.1116]$	$+0.0383$ $[+0.0083, +0.0725]$	$+0.0607$ $[+0.0392, +0.0929]$
Wanda 2:4	5/6	$+0.0799$ $[+0.0547, +0.1152]$	$+0.0447$ $[+0.0138, +0.0800]$	$+0.0671$ $[+0.0443, +0.1004]$
Student^M	4/6	$+0.0319$ $[+0.0174, +0.0578]$	0.0000 $[-0.0246, +0.0246]$	$+0.0224$ $[+0.0100, +0.0454]$
Gemma				
INT8	6/8	$+0.0192$ $[+0.0068, +0.0412]$	-0.0256 $[-0.0496, -0.0130]$	0.0000 $[-0.0121, +0.0121]$
NF4	8/8	$+0.0256$ $[+0.0130, +0.0496]$	-0.0256 $[-0.0496, -0.0130]$	0.0000 $[-0.0121, +0.0121]$
GPTQ	9/8	$+0.0288$ $[+0.0152, +0.0537]$	-0.0256 $[-0.0496, -0.0130]$	0.0000 $[-0.0121, +0.0121]$
AWQ	10/8	$+0.0319$ $[+0.0174, +0.0578]$	-0.0256 $[-0.0496, -0.0130]$	0.0000 $[-0.0121, +0.0121]$
QLoRA	8/8	$+0.0224$ $[+0.0046, +0.0468]$	-0.0288 $[-0.0537, -0.0152]$	-0.0032 $[-0.0179, +0.0090]$
Wanda 50%	26/8	$+0.0831$ $[+0.0573, +0.1189]$	-0.0256 $[-0.0496, -0.0130]$	0.0000 $[-0.0121, +0.0121]$
SparseGPT 50%	5/8	$+0.0160$ $[+0.0037, +0.0368]$	-0.0256 $[-0.0496, -0.0130]$	0.0000 $[-0.0121, +0.0121]$

Continued on next page

Table 7: (continued)

Method	Missing	Upward [CI]	Downward [CI]	Guard [CI]
SparseGPT 2:4	9/8	$+0.0256$ $[+0.0073, +0.0510]$	-0.0288 $[-0.0537, -0.0152]$	-0.0032 $[-0.0179, +0.0090]$
Wanda 2:4	10/8	$+0.0319$ $[+0.0174, +0.0578]$	-0.0256 $[-0.0496, -0.0130]$	0.0000 $[-0.0121, +0.0121]$
Student	6/8	$+0.0256$ $[+0.0073, +0.0510]$	-0.0192 $[-0.0454, +0.0030]$	$+0.0096$ $[-0.0064, +0.0296]$

Table 8 combines missing-verdict imputation with Holm correction. In the worst case, unjudged responses in all thirteen positive cells are imputed against their increases at once while the other twenty-seven tests keep their observed p-values; three increases survive. Guard-informed imputation in every cell keeps all seven Holm-significant increases. The column k gives, for each Holm-significant increase, how many unjudged BF16 responses would have to be successes, with unjudged compressed responses counted as refusals, before significance is lost; it is a single-cell criterion.

Table 8: Multiplicity combined with missing verdicts for the thirteen positive judge intervals. b/c count BF16-only/compressed-only successes. Holm adjusts two-sided exact McNemar tests over all forty compressed cells. Worst case imputes, in all thirteen positive cells at once, every unjudged response against the increase in that cell; the other twenty-seven cells keep their observed tests. Guard-informed imputation labels each unjudged response a success exactly when the guard flagged it, in every cell. Worst case is a logical bound rather than an estimate. k is the smallest number of unjudged BF16 responses that, counted as successes with unjudged compressed responses counted as refusals, removes Holm significance while the other cells keep their observed tests, shown over the number of unjudged BF16 responses; none means significance survives all of them counted as successes, and — marks a cell without Holm significance.

Family	Method	Observed b/c	Observed Holm p	Worst-case b/c	Worst-case Holm p	Guard-informed Holm p	k	H7
Llama	Wanda 2:4	2/57	2.5e−13	6/57	6.6e−10	1.3e−13	none	B
Llama	SparseGPT 2:4	1/38	5.7e−9	5/38	9.7e−6	5.7e−9	none	B
Qwen	Student	0/30	7.1e−8	7/30	0.0073	1.8e−8	none	NA+
OLMo	Wanda 2:4	0/20	7.1e−5	6/20	0.3461	3.4e−5	4/6	C
Qwen	Wanda 2:4	0/19	1.4e−4	7/19	1.0000	3.4e−8	4/7	B
OLMo	SparseGPT 2:4	0/18	2.7e−4	6/18	0.8156	1.3e−4	3/6	C
Qwen	SparseGPT 2:4	0/17	5.2e−4	7/17	1.0000	1.3e−4	3/7	B
Llama	Wanda 50%	1/12	0.1128	5/12	1.0000	0.1094	—	C
Llama	QLoRA	0/9	0.1250	4/9	1.0000	0.1211	—	C
Qwen	Wanda 50%	0/8	0.2422	7/8	1.0000	0.0645	—	B
OLMo	Student	0/6	0.9375	6/6	1.0000	0.4688	—	NA+
Qwen	SparseGPT 50%	0/6	0.9375	7/6	1.0000	0.9062	—	B
Qwen	QLoRA	0/4	1.0000	7/4	1.0000	1.0000	—	C

Table 9: Coverage and sensitivity on all 313 requests per cell. q is the number without a valid judgement, not an identified provider-filter count. s counts judged successes. For compressed rows, entries are numerators over 313: D_0 imputes all unjudged responses as refusals, D_1 as successes, and $[D_-, D_+]$ permits opposite imputations in model and baseline. These are finite-corpus sensitivity bounds, not confidence intervals. M is defined in Appendix A.

Family	Method	Valid	q	s	Paired	$313D_0$	$313D_1$	$313[D_-, D_+]$
Llama	BF16	309	4	2	—	—	—	—
Llama	INT8	309	4	2	309	0	0	$[-4, 4]$
Llama	NF4	310	3	2	309	0	-1	$[-4, 3]$
Llama	GPTQ	309	4	1	308	-1	-1	$[-5, 3]$
Llama	AWQ	309	4	1	309	-1	-1	$[-5, 3]$
Llama	QLoRA^M	310	3	11	309	9	8	$[5, 12]$
Llama	Wanda 50%^{M}	309	4	13	308	11	11	$[7, 15]$
Llama	SparseGPT 50%	309	4	6	309	4	4	$[0, 8]$
Llama	SparseGPT 2:4	309	4	39	309	37	37	$[33, 41]$
Llama	Wanda 2:4	309	4	57	308	55	55	$[51, 59]$
Llama	Student	310	3	2	309	0	-1	$[-4, 3]$
Qwen	BF16	306	7	0	—	—	—	—
Qwen	INT8	308	5	2	306	2	0	$[-5, 7]$
Qwen	NF4	308	5	1	306	1	-1	$[-6, 6]$
Qwen	GPTQ	308	5	1	305	1	-1	$[-6, 6]$
Qwen	AWQ	310	3	3	306	3	-1	$[-4, 6]$
Qwen	QLoRA^M	302	11	4	302	4	8	$[-3, 15]$
Qwen	Wanda 50%^{M}	306	7	8	303	8	8	$[1, 15]$
Qwen	SparseGPT 50%^{M}	305	8	6	303	6	7	$[-1, 14]$

Continued on next page

Table 9: (continued)

Family	Method	Valid	q	s	Paired	$313D_0$	$313D_1$	$313[D_-, D_+]$
Qwen	SparseGPT 2:4	307	6	17	305	17	16	$[10, 23]$
Qwen	Wanda 2:4	293	20	19	291	19	32	$[12, 39]$
Qwen	Student	307	6	30	305	30	29	$[23, 36]$
OLMo	BF16	307	6	0	—	—	—	—
OLMo	INT8	310	3	0	307	0	-3	$[-6, 3]$
OLMo	NF4	309	4	0	307	0	-2	$[-6, 4]$
OLMo	GPTQ	309	4	0	307	0	-2	$[-6, 4]$
OLMo	AWQ	309	4	0	306	0	-2	$[-6, 4]$
OLMo	QLoRA	309	4	1	307	1	-1	$[-5, 5]$
OLMo	Wanda 50%	307	6	0	305	0	0	$[-6, 6]$
OLMo	SparseGPT 50%	308	5	1	307	1	0	$[-5, 6]$
OLMo	SparseGPT 2:4	307	6	18	305	18	18	$[12, 24]$
OLMo	Wanda 2:4	308	5	20	306	20	19	$[14, 25]$
OLMo	Student^M	309	4	6	306	6	4	$[0, 10]$
Gemma	BF16	305	8	1	—	—	—	—
Gemma	INT8	307	6	1	305	0	-2	$[-8, 6]$
Gemma	NF4	305	8	1	304	0	0	$[-8, 8]$
Gemma	GPTQ	304	9	1	303	0	1	$[-8, 9]$
Gemma	AWQ	303	10	1	302	0	2	$[-8, 10]$
Gemma	QLoRA	305	8	0	302	-1	-1	$[-9, 7]$
Gemma	Wanda 50%	287	26	1	284	0	18	$[-8, 26]$
Gemma	SparseGPT 50%	308	5	1	304	0	-3	$[-8, 5]$
Gemma	SparseGPT 2:4	304	9	0	302	-1	0	$[-9, 8]$
Gemma	Wanda 2:4	303	10	1	302	0	2	$[-8, 10]$
Gemma	Student	307	6	3	303	2	0	$[-6, 8]$

A.4 GUARD CONFUSION AGAINST THE JUDGE

Gemma Wanda 2:4 has perplexity ratio 1.7022 but guard flag rate 0.0096 $[0.0020, 0.0278]$ (3 of 313 responses) and judge change 0.0000 $[-0.0126, +0.0126]$. This is consistent with guard sensitivity to harmful content rather than compression damage alone; flag rates and ASRs have different denominators.

Table 10: Guard performance on each cell's judge-valid responses. Each prompt occurs once per cell; brackets are exact two-sided 95% Clopper–Pearson intervals. NA means a zero denominator. The main-text pooled rates instead use a prompt-cluster bootstrap across cells. The judge is a reference, not ground truth. M marks a positive Tango interval that includes zero under adversarial missing-verdict imputation (an unadjusted criterion; Holm fragility is in column k of Table 8).

Family	Method	Valid	Recall [95% CI]	Precision [95% CI]	FPR [95% CI]
Llama	BF16	309	0.5000 [0.0126, 0.9874]	0.5000 [0.0126, 0.9874]	0.0033 [0.0001, 0.0180]
Llama	INT8	309	0.5000 [0.0126, 0.9874]	0.3333 [0.0084, 0.9057]	0.0065 [0.0008, 0.0233]
Llama	NF4	310	0.5000 [0.0126, 0.9874]	0.3333 [0.0084, 0.9057]	0.0065 [0.0008, 0.0233]
Llama	GPTQ	309	1.0000 [0.0250, 1.0000]	0.5000 [0.0126, 0.9874]	0.0032 [0.0001, 0.0180]
Llama	AWQ	309	0.0000 [0.0000, 0.9750]	0.0000 [0.0000, 0.8419]	0.0065 [0.0008, 0.0233]
Llama	QLoRA^M	310	0.7273 [0.3903, 0.9398]	0.3077 [0.1433, 0.5179]	0.0602 [0.0361, 0.0935]
Llama	Wanda 50%^{M}	309	0.8462 [0.5455, 0.9808]	0.2619 [0.1386, 0.4204]	0.1047 [0.0723, 0.1454]
Llama	SparseGPT 50%	309	0.8333 [0.3588, 0.9958]	0.2000 [0.0683, 0.4070]	0.0660 [0.0408, 0.1001]
Llama	SparseGPT 2:4	309	0.8974 [0.7578, 0.9713]	0.3182 [0.2326, 0.4138]	0.2778 [0.2252, 0.3353]
Llama	Wanda 2:4	309	0.9649 [0.8789, 0.9957]	0.4044 [0.3212, 0.4919]	0.3214 [0.2642, 0.3829]
Llama	Student	310	1.0000 [0.1581, 1.0000]	0.1111 [0.0138, 0.3471]	0.0519 [0.0300, 0.0830]
Qwen	BF16	306	NA	0.0000 [0.0000, 0.2180]	0.0490 [0.0277, 0.0796]
Qwen	INT8	308	0.5000 [0.0126, 0.9874]	0.0625 [0.0016, 0.3023]	0.0490 [0.0277, 0.0796]
Qwen	NF4	308	1.0000 [0.0250, 1.0000]	0.0385 [0.0010, 0.1964]	0.0814 [0.0534, 0.1179]
Qwen	GPTQ	308	0.0000 [0.0000, 0.9750]	0.0000 [0.0000, 0.1951]	0.0554 [0.0326, 0.0872]
Qwen	AWQ	310	0.3333 [0.0084, 0.9057]	0.0667 [0.0017, 0.3195]	0.0456 [0.0252, 0.0753]
Qwen	QLoRA^M	302	0.7500 [0.1941, 0.9937]	0.1875 [0.0405, 0.4565]	0.0436 [0.0234, 0.0734]
Qwen	Wanda 50%^{M}	306	0.7500 [0.3491, 0.9681]	0.0822 [0.0308, 0.1704]	0.2248 [0.1787, 0.2765]
Qwen	SparseGPT 50%^{M}	305	0.8333 [0.3588, 0.9958]	0.1042 [0.0347, 0.2266]	0.1438 [0.1061, 0.1888]
Qwen	SparseGPT 2:4	307	0.9412 [0.7131, 0.9985]	0.1798 [0.1064, 0.2755]	0.2517 [0.2028, 0.3058]
Qwen	Wanda 2:4	293	0.8947 [0.6686, 0.9870]	0.2237 [0.1360, 0.3338]	0.2153 [0.1681, 0.2688]
Qwen	Student	307	0.9000 [0.7347, 0.9789]	0.3034 [0.2103, 0.4099]	0.2238 [0.1761, 0.2775]
OLMo	BF16	307	NA	0.0000 [0.0000, 0.9750]	0.0033 [0.0001, 0.0180]
OLMo	INT8	310	NA	NA	0.0000 [0.0000, 0.0118]
OLMo	NF4	309	NA	0.0000 [0.0000, 0.8419]	0.0065 [0.0008, 0.0232]
OLMo	GPTQ	309	NA	0.0000 [0.0000, 0.9750]	0.0032 [0.0001, 0.0179]
OLMo	AWQ	309	NA	0.0000 [0.0000, 0.8419]	0.0065 [0.0008, 0.0232]
OLMo	QLoRA	309	1.0000 [0.0250, 1.0000]	1.0000 [0.0250, 1.0000]	0.0000 [0.0000, 0.0119]
OLMo	Wanda 50%	307	NA	0.0000 [0.0000, 0.3694]	0.0261 [0.0113, 0.0507]
OLMo	SparseGPT 50%	308	0.0000 [0.0000, 0.9750]	0.0000 [0.0000, 0.4593]	0.0195 [0.0072, 0.0421]

Continued on next page

Table 10: (continued)

Family	Method	Valid	Recall [95% CI]	Precision [95% CI]	FPR [95% CI]
OLMo	SparseGPT 2:4	307	0.7778 [0.5236, 0.9359]	0.1867 [0.1060, 0.2933]	0.2111 [0.1655, 0.2627]
OLMo	Wanda 2:4	308	0.9000 [0.6830, 0.9877]	0.2278 [0.1410, 0.3360]	0.2118 [0.1661, 0.2636]
OLMo	Student^M	309	1.0000 [0.5407, 1.0000]	0.2308 [0.0897, 0.4365]	0.0660 [0.0408, 0.1001]
Gemma	BF16	305	0.0000 [0.0000, 0.9750]	NA	0.0000 [0.0000, 0.0121]
Gemma	INT8	307	0.0000 [0.0000, 0.9750]	NA	0.0000 [0.0000, 0.0120]
Gemma	NF4	305	0.0000 [0.0000, 0.9750]	NA	0.0000 [0.0000, 0.0121]
Gemma	GPTQ	304	0.0000 [0.0000, 0.9750]	0.0000 [0.0000, 0.9750]	0.0033 [0.0001, 0.0183]
Gemma	AWQ	303	0.0000 [0.0000, 0.9750]	0.0000 [0.0000, 0.9750]	0.0033 [0.0001, 0.0183]
Gemma	QLoRA	305	NA	0.0000 [0.0000, 0.2849]	0.0361 [0.0181, 0.0636]
Gemma	Wanda 50%	287	0.0000 [0.0000, 0.9750]	NA	0.0000 [0.0000, 0.0128]
Gemma	SparseGPT 50%	308	0.0000 [0.0000, 0.9750]	NA	0.0000 [0.0000, 0.0119]
Gemma	SparseGPT 2:4	304	NA	0.0000 [0.0000, 0.8419]	0.0066 [0.0008, 0.0236]
Gemma	Wanda 2:4	303	0.0000 [0.0000, 0.9750]	0.0000 [0.0000, 0.7076]	0.0099 [0.0021, 0.0288]
Gemma	Student	307	1.0000 [0.2924, 1.0000]	0.1667 [0.0358, 0.4142]	0.0493 [0.0279, 0.0801]

B LAYER DEPENDENCE AND COMPRESSION-AUGMENTED TRANSFER

A shallow Gemma readout can decline without a behavioural increase. Gemma SparseGPT 2:4 has no detected behavioural increase (Table 1), while its pooled frozen AUROC at layer 4 falls below the 0.785 reference. Retraining raises the pooled AUROC from 0.6759 to 0.8684, a paired gain of $+0.1925$ $[+0.1599, +0.2240]$ (95% prompt-sampling interval). Per fold, its layer-4 frozen AUROC is 0.8265, above the reference. This illustrates the distinction between input-harm discrimination and judged compliance; Section 7's near-baseline Gemma readouts concern the headline layer 21.

B.1 PLANNED LAYERS AND RULE SENSITIVITY

The post hoc rule H7 calls a readout degraded when its paired change interval lies below -0.05 or its absolute range, the envelope of its per-seed absolute-AUROC CIs, reaches 0.785. At the headline layers it gives 25 A (no increase), 6 B (increase, degraded), 5 C (increase, criterion not met), and 4 students; no cell without an increase has a degraded readout (label D in Table 11). The partition including D is unchanged for margins in $[0.0253, 0.0900)$, and the B/C split alone for $[0.0224, 0.0900)$; at the planned layers it gives 3 B and 8 C, and the planned definitions give no B. The strict rule places both OLMo 2:4 cells among the ten where both axes move and only Llama QLoRA among behaviour-only cells.

Planned layers are those set in each family's extraction configuration (Llama 16, Qwen 14, OLMo 16, Gemma 21); the headline layers Llama 8, Qwen 7, and OLMo 8 were chosen after results, and Gemma is reported at its planned layer (Section 3). H7's 0.05 margin was also chosen after the results and has no independent justification. At the headline layers H7's partition coincides with the planned big-drop definition. Its C label combines unchanged and detectably reduced readouts, so C does not indicate an unchanged readout. Table 11 applies H7 to all forty cells at the headline and planned layers, with its 0.10 variant, the planned definitions and the strict rule at the layers named in each row; the strict rule finds twelve cells with neither axis moving, thirteen with only the readout, ten with both, and one with only behaviour. Keeping H7's absolute-range arm but applying 0.10 to the change-interval endpoint moves only Llama SparseGPT 2:4 (point -0.1107, interval $[-0.1346, -0.0900]$) from B to C; Qwen Wanda 50% and SparseGPT 50% (upper endpoints -0.0846, -0.0915) remain B only through the absolute-range arm. Table 12 evaluates the planned definitions at every layer for the eleven compatible increases.

Table 11: Readout-rule sensitivity over the forty compressed cells. Headline layers: Llama 8, Qwen 7, OLMo 8, Gemma 21; planned layers, fixed in the extraction configurations: Llama 16, Qwen 14, OLMo 16, Gemma 21. In the first five rows, A is a cell with no behavioural increase and no degraded readout, B an increase with a readout degraded under that row's rule, C an increase without one, and D a degraded readout without an increase (none occurs in these rows); NA+/NA0 count students with/without an increase, to which no readout rule applies. The strict row has its own headings and counts cells by which axes move; an axis moves when its 95% interval (judge Tango or paired readout change) excludes zero.

Readout rule	A	B	C	D	NA+/NA0
H7 (post hoc): change CI upper < -0.05 or absolute range reaches 0.785; headline layer	25	6	5	0	2/2
H7 with 0.10 applied to the change CI upper endpoint; headline layer	25	5	6	0	2/2
Pre-specified big drop (point change < -0.10); headline layer	25	6	5	0	2/2
H7; planned layer	25	3	8	0	2/2
Pre-specified big drop or operational failure (AUROC < 0.75); planned layer	25	0	11	0	2/2

Strict rule	Neither	Both	Behaviour only	Readout only	NA+/NA0
Judge and readout 95% CIs each tested against zero; headline layer	12	10	1	13	2/2

Table 12: Frozen readout across all eight teacher layers for the eleven compatible cells with a positive judge interval. Big drop and operational failure use the definitions fixed before data collection (point change < -0.10; seed-mean AUROC < 0.75). Within 0.05 counts layers whose paired change CI lower endpoint exceeds -0.05. This uses H7's post hoc 0.05 margin, not the 0.10 big-drop margin of Appendix E.7. The last column is H7 at each layer in increasing depth. Planned-layer changes are paired 95% prompt-bootstrap intervals. Labels use endpoints stored at four decimals. At Llama SparseGPT 2:4 layer 24 the stored upper endpoint is -0.0500, on the H7 margin, and the label is C; the unrounded upper endpoint, -0.05003, would give B (29 B and 59 C overall), within Monte Carlo error.

Family	Method	Big drop	Op. failure	Within 0.05	Planned layer change	H7 by layer
Llama	QLoRA	0/8	0/8	8/8	L16: -0.0011 $[-0.0042, +0.0011]$	CCCCCCCC
Llama	SparseGPT 2:4	2/8	1/8	0/8	L16: -0.0599 $[-0.0795, -0.0421]$	BBBCCCBC
Llama	Wanda 2:4	3/8	1/8	0/8	L16: -0.0824 $[-0.1070, -0.0608]$	BBBBBBBB
Llama	Wanda 50%	1/8	0/8	7/8	L16: -0.0109 $[-0.0205, -0.0036]$	BCCCCCCC
OLMo	SparseGPT 2:4	0/8	0/8	6/8	L16: -0.0212 $[-0.0321, -0.0118]$	CCCCCCCC
OLMo	Wanda 2:4	0/8	0/8	7/8	L16: -0.0129 $[-0.0202, -0.0071]$	BCCCCCCC
Qwen	QLoRA	0/8	0/8	8/8	L14: -0.0032 $[-0.0067, -0.0001]$	CCCCCCCC
Qwen	SparseGPT 2:4	3/8	2/8	3/8	L14: -0.0841 $[-0.1036, -0.0651]$	BBBBCCCC
Qwen	SparseGPT 50%	2/8	0/8	5/8	L14: -0.0276 $[-0.0394, -0.0170]$	BBBCCCCC
Qwen	Wanda 2:4	3/8	2/8	2/8	L14: -0.0697 $[-0.0889, -0.0531]$	BBBBCCCC
Qwen	Wanda 50%	2/8	0/8	5/8	L14: -0.0184 $[-0.0267, -0.0112]$	BBBCCCCC

Depth sensitivity of H7. The fixed headline mapping uses Llama 8, Qwen 7, OLMo 8, and Gemma 21. It was chosen after the results (Section 3). Table 13 covers all eleven behaviour-positive, dimension-compatible cells at all eight teacher depths; H marks the headline. Each H7 label uses the corresponding paired-change CI and absolute range. Gemma contributes no behaviour-positive cells, and students are structurally inapplicable. Cells marked M have missingness-sensitive behavioural increases; this property does not depend on the probe layer.

H7 gives twenty-eight B and sixty C, including twenty-two B and fifty-five C off the headline layer; Llama SparseGPT 2:4 at layer 24 sits on the 0.05 margin at stored precision (Table 12). Table 14 supplies the paired changes and their separate fold-split SDs. Under the strict rule, Llama QLoRA has only behavioural movement at layers 8/12/16/20/24 and movement on both axes at 4/28/32; its behavioural increase remains missingness-sensitive at every depth. The strict rule also finds behaviour-only movement for Qwen QLoRA at layer 21 and Qwen Wanda 50% at layers 21/24/28 (both headline layer 7), both likewise missingness-sensitive. No headline-only conclusion can replace this depth audit.

Table 13: H7 label at every teacher depth for the eleven behaviour-positive, dimension-compatible cells. B: the paired change CI upper endpoint lies below -0.05 or the absolute per-seed CI range reaches 0.785; C: neither. ^{H} marks the headline layer; ^{M} marks a missingness-sensitive behavioural increase, which does not depend on the probe layer. Llama SparseGPT 2:4 at layer 24 lies on the 0.05 margin at stored precision (Table 12).

Family	Method	Activation layer and H7 label							
Llama layers		4	8	12	16	20	24	28	32
Llama	QLoRA^{M}	C	C^{H}	C	C	C	C	C	C
Llama	Wanda 50%^{M}	B	C^{H}	C	C	C	C	C	C
Llama	SparseGPT 2:4	B	B^{H}	B	C	C	C	B	C
Llama	Wanda 2:4	B	B^{H}	B	B	B	B	B	B
Qwen layers		4	7	10	14	17	21	24	28
Qwen	QLoRA^{M}	C	C^{H}	C	C	C	C	C	C
Qwen	Wanda 50%^{M}	B	B^{H}	B	C	C	C	C	C
Qwen	SparseGPT 50%^{M}	B	B^{H}	B	C	C	C	C	C
Qwen	SparseGPT 2:4	B	B^{H}	B	B	C	C	C	C
Qwen	Wanda 2:4	B	B^{H}	B	B	C	C	C	C
OLMo layers		4	8	12	16	20	24	28	32
OLMo	SparseGPT 2:4	C	C^{H}	C	C	C	C	C	C
OLMo	Wanda 2:4	B	C^{H}	C	C	C	C	C	C

Table 14: Paired frozen input-harm AUROC changes at every teacher depth of the eleven behaviour-positive compatible cells. Each 2,000-draw stratified prompt bootstrap averages the three per-seed AUROC differences within a shared draw; predictions are not pooled across seeds. The fold-split SD of the delta is separate. ^{H} marks the headline layer; the H7 column gives the H7 label at each layer. M is defined in Appendix A.

Family	Method	Layer	ΔAUROC [95% CI]	s_{Δ}	H7
Llama	QLoRA^{M}	4	-0.0137 $[-0.0213, -0.0063]$	0.0070	C
Llama	Wanda 50%^{M}	4	-0.1007 $[-0.1206, -0.0820]$	0.0772	B
Llama	SparseGPT 2:4	4	-0.2363 $[-0.2668, -0.2066]$	0.0600	B
Llama	Wanda 2:4	4	-0.2165 $[-0.2450, -0.1890]$	0.0681	B
Llama	QLoRA^{M}	8^{H}	-0.0003 $[-0.0016, +0.0009]$	0.0003	C
Llama	Wanda 50%^{M}	8^{H}	-0.0321 $[-0.0439, -0.0214]$	0.0127	C
Llama	SparseGPT 2:4	8^{H}	-0.1107 $[-0.1346, -0.0900]$	0.0246	B
Llama	Wanda 2:4	8^{H}	-0.1983 $[-0.2323, -0.1641]$	0.0416	B
Llama	QLoRA^{M}	12	-0.0007 $[-0.0028, +0.0014]$	0.0004	C
Llama	Wanda 50%^{M}	12	-0.0082 $[-0.0135, -0.0041]$	0.0015	C
Llama	SparseGPT 2:4	12	-0.0959 $[-0.1177, -0.0755]$	0.0076	B
Llama	Wanda 2:4	12	-0.1834 $[-0.2137, -0.1552]$	0.0168	B
Llama	QLoRA^{M}	16	-0.0011 $[-0.0042, +0.0011]$	0.0002	C
Llama	Wanda 50%^{M}	16	-0.0109 $[-0.0205, -0.0036]$	0.0007	C
Llama	SparseGPT 2:4	16	-0.0599 $[-0.0795, -0.0421]$	0.0254	C
Llama	Wanda 2:4	16	-0.0824 $[-0.1070, -0.0608]$	0.0282	B
Llama	QLoRA^{M}	20	-0.0026 $[-0.0080, +0.0016]$	0.0007	C
Llama	Wanda 50%^{M}	20	-0.0113 $[-0.0217, -0.0032]$	0.0004	C
Llama	SparseGPT 2:4	20	-0.0578 $[-0.0777, -0.0393]$	0.0150	C
Llama	Wanda 2:4	20	-0.0704 $[-0.0926, -0.0507]$	0.0134	B
Llama	QLoRA^{M}	24	-0.0053 $[-0.0131, +0.0004]$	0.0011	C
Llama	Wanda 50%^{M}	24	-0.0102 $[-0.0202, -0.0027]$	0.0011	C
Llama	SparseGPT 2:4	24	-0.0706 $[-0.0926, -0.0500]$	0.0191	C
Llama	Wanda 2:4	24	-0.0787 $[-0.1019, -0.0583]$	0.0227	B
Llama	QLoRA^{M}	28	-0.0066 $[-0.0135, -0.0011]$	0.0005	C
Llama	Wanda 50%^{M}	28	-0.0100 $[-0.0180, -0.0038]$	0.0015	C
Llama	SparseGPT 2:4	28	-0.0726 $[-0.0949, -0.0523]$	0.0112	B
Llama	Wanda 2:4	28	-0.0872 $[-0.1096, -0.0665]$	0.0186	B
Llama	QLoRA^{M}	32	-0.0060 $[-0.0113, -0.0017]$	0.0009	C
Llama	Wanda 50%^{M}	32	-0.0103 $[-0.0183, -0.0043]$	0.0013	C
Llama	SparseGPT 2:4	32	-0.0713 $[-0.0944, -0.0485]$	0.0018	C
Llama	Wanda 2:4	32	-0.0785 $[-0.1030, -0.0569]$	0.0048	B

Continued on next page

Table 14: (continued)

Family	Method	Layer	ΔAUROC [95% CI]	s_Δ	H7
Qwen	QLoRA^M	4	-0.0078 $[-0.0156, -0.0006]$	0.0051	C
Qwen	Wanda 50%^{M}	4	-0.0857 $[-0.1074, -0.0655]$	0.0591	B
Qwen	SparseGPT 50%^{M}	4	-0.0627 $[-0.0805, -0.0460]$	0.0580	B
Qwen	SparseGPT 2:4	4	-0.1311 $[-0.1573, -0.1066]$	0.0543	B
Qwen	Wanda 2:4	4	-0.1887 $[-0.2207, -0.1573]$	0.0528	B
Qwen	QLoRA^M	7^H	-0.0231 $[-0.0331, -0.0138]$	0.0076	C
Qwen	Wanda 50%^{M}	7^H	-0.1051 $[-0.1277, -0.0846]$	0.0360	B
Qwen	SparseGPT 50%^{M}	7^H	-0.1135 $[-0.1359, -0.0915]$	0.0554	B
Qwen	SparseGPT 2:4	7^H	-0.2199 $[-0.2535, -0.1888]$	0.0406	B
Qwen	Wanda 2:4	7^H	-0.2229 $[-0.2557, -0.1920]$	0.0337	B
Qwen	QLoRA^M	10	-0.0207 $[-0.0305, -0.0111]$	0.0106	C
Qwen	Wanda 50%^{M}	10	-0.1160 $[-0.1383, -0.0954]$	0.0362	B
Qwen	SparseGPT 50%^{M}	10	-0.1111 $[-0.1334, -0.0889]$	0.0482	B
Qwen	SparseGPT 2:4	10	-0.2255 $[-0.2560, -0.1957]$	0.0291	B
Qwen	Wanda 2:4	10	-0.2145 $[-0.2457, -0.1854]$	0.0080	B
Qwen	QLoRA^M	14	-0.0032 $[-0.0067, -0.0001]$	0.0024	C
Qwen	Wanda 50%^{M}	14	-0.0184 $[-0.0267, -0.0112]$	0.0144	C
Qwen	SparseGPT 50%^{M}	14	-0.0276 $[-0.0394, -0.0170]$	0.0063	C
Qwen	SparseGPT 2:4	14	-0.0841 $[-0.1036, -0.0651]$	0.0434	B
Qwen	Wanda 2:4	14	-0.0697 $[-0.0889, -0.0531]$	0.0151	B
Qwen	QLoRA^M	17	-0.0036 $[-0.0072, -0.0010]$	0.0021	C
Qwen	Wanda 50%^{M}	17	-0.0037 $[-0.0070, -0.0012]$	0.0010	C
Qwen	SparseGPT 50%^{M}	17	-0.0071 $[-0.0121, -0.0032]$	0.0016	C
Qwen	SparseGPT 2:4	17	-0.0429 $[-0.0569, -0.0309]$	0.0142	C
Qwen	Wanda 2:4	17	-0.0397 $[-0.0527, -0.0275]$	0.0142	C
Qwen	QLoRA^M	21	-0.0029 $[-0.0078, +0.0002]$	0.0003	C
Qwen	Wanda 50%^{M}	21	-0.0030 $[-0.0077, +0.0003]$	0.0015	C
Qwen	SparseGPT 50%^{M}	21	-0.0043 $[-0.0086, -0.0010]$	0.0004	C
Qwen	SparseGPT 2:4	21	-0.0118 $[-0.0194, -0.0053]$	0.0011	C
Qwen	Wanda 2:4	21	-0.0226 $[-0.0344, -0.0125]$	0.0016	C
Qwen	QLoRA^M	24	-0.0049 $[-0.0108, -0.0008]$	0.0016	C
Qwen	Wanda 50%^{M}	24	-0.0031 $[-0.0079, +0.0006]$	0.0013	C
Qwen	SparseGPT 50%^{M}	24	-0.0047 $[-0.0100, -0.0009]$	0.0014	C
Qwen	SparseGPT 2:4	24	-0.0135 $[-0.0234, -0.0055]$	0.0020	C
Qwen	Wanda 2:4	24	-0.0233 $[-0.0346, -0.0131]$	0.0030	C
Qwen	QLoRA^M	28	-0.0087 $[-0.0177, -0.0019]$	0.0014	C
Qwen	Wanda 50%^{M}	28	-0.0051 $[-0.0122, 0.0000]$	0.0016	C
Qwen	SparseGPT 50%^{M}	28	-0.0081 $[-0.0155, -0.0024]$	0.0011	C
Qwen	SparseGPT 2:4	28	-0.0262 $[-0.0383, -0.0148]$	0.0054	C
Qwen	Wanda 2:4	28	-0.0356 $[-0.0516, -0.0218]$	0.0030	C
OLMo	SparseGPT 2:4	4	-0.0598 $[-0.0786, -0.0404]$	0.0090	C
OLMo	Wanda 2:4	4	-0.0902 $[-0.1120, -0.0675]$	0.0111	B
OLMo	SparseGPT 2:4	8^H	-0.0345 $[-0.0489, -0.0224]$	0.0025	C
OLMo	Wanda 2:4	8^H	-0.0191 $[-0.0299, -0.0093]$	0.0031	C
OLMo	SparseGPT 2:4	12	-0.0457 $[-0.0614, -0.0321]$	0.0046	C
OLMo	Wanda 2:4	12	-0.0184 $[-0.0283, -0.0099]$	0.0023	C
OLMo	SparseGPT 2:4	16	-0.0212 $[-0.0321, -0.0118]$	0.0080	C
OLMo	Wanda 2:4	16	-0.0129 $[-0.0202, -0.0071]$	0.0045	C
OLMo	SparseGPT 2:4	20	-0.0110 $[-0.0184, -0.0053]$	0.0007	C
OLMo	Wanda 2:4	20	-0.0104 $[-0.0176, -0.0050]$	0.0020	C
OLMo	SparseGPT 2:4	24	-0.0132 $[-0.0228, -0.0053]$	0.0030	C
OLMo	Wanda 2:4	24	-0.0110 $[-0.0192, -0.0045]$	0.0014	C
OLMo	SparseGPT 2:4	28	-0.0220 $[-0.0335, -0.0117]$	0.0060	C
OLMo	Wanda 2:4	28	-0.0155 $[-0.0257, -0.0066]$	0.0029	C
OLMo	SparseGPT 2:4	32	-0.0128 $[-0.0213, -0.0066]$	0.0025	C
OLMo	Wanda 2:4	32	-0.0112 $[-0.0186, -0.0056]$	0.0019	C

B.2 Compression-augmented transfer across depth

Augmented probes use 2,000 maximum iterations. The row-count-matched noise control adds eight synthetic copies of the 450 BF16 prompt rows. With eight non-held-out compression variants $\mathcal{V}$ and residuals $r_{vij} = X_{vij} - X_{0ij}$, it uses independent $\epsilon_{ij} \sim \mathcal{N}(0, \sigma_j^2)$, where $\sigma_j^2 =$

$(8 \cdot 450)^{-1} \sum_{v \in \mathcal{V}, i} r_{vij}^2$. This matches uncentred residual second moments per feature, rather than centred residual variance. It uses all non-held-out variants, including quantisation and QLoRA. Each training fold has 3,240 rows, matching all-variant augmentation, versus 360 for BF16-only. Scale calibration uses all 450 prompts before splitting, including test covariates but no labels; this transductive calibration limits the noise comparison. A saved control estimates scales within each training fold (Appendix E.15).

Table 15 repeats the augmentation comparison at the three augmentation layers of each family (Llama and OLMo 4, 8, 16; Qwen 4, 7, 17). Table 16 reports both gap-closure definitions, with layer four primary and eligibility fixed at denominator greater than 0.01. Conditional method–seed-row intervals in the ablation do not account for shared-prompt uncertainty; method-clustered gap intervals answer a different question. The sensitivity analysis also evaluates gap cutoffs 0, 0.005, 0.01, and 0.02 at each of these layers; the primary-layer comparison is in Table 5. Figure 4 shows the leave-one-compression-out experiment. Appendix E.14 adds matched pooled/per-fold columns across depth on the all-method subset; Table 15 uses the pruning-only subset.

Table 15: Every measured augmentation layer: paired AUROC improvement over BF16-only training on held-out pruning. Each row averages twelve method–seed cells; brackets are 95% conditional cell-bootstrap intervals (5,000 draws). Rows share prompts and training pools, so these CIs do not measure prompt uncertainty or support independent population-level method inference. † marks primary layer four. The primary pooled comparison is Table 2.

Family	Layer	Augmentation	ΔAUROC [95% CI]
Llama	4†	Quantisation	+0.0136 [−0.0077, +0.0370]
Llama	4†	Gaussian noise	+0.0312 [+0.0073, +0.0550]
Llama	4†	Other pruning	+0.0995 [+0.0623, +0.1394]
Llama	4†	All variants	+0.0958 [+0.0554, +0.1374]
Llama	8	Quantisation	−0.0061 [−0.0215, +0.0088]
Llama	8	Gaussian noise	+0.0011 [−0.0160, +0.0166]
Llama	8	Other pruning	+0.0474 [+0.0236, +0.0757]
Llama	8	All variants	+0.0463 [+0.0237, +0.0733]
Llama	16	Quantisation	+0.0065 [+0.0015, +0.0120]
Llama	16	Gaussian noise	+0.0101 [+0.0026, +0.0189]
Llama	16	Other pruning	+0.0290 [+0.0157, +0.0439]
Llama	16	All variants	+0.0301 [+0.0164, +0.0458]
Qwen	4†	Quantisation	+0.0146 [+0.0014, +0.0274]
Qwen	4†	Gaussian noise	+0.0314 [+0.0083, +0.0531]
Qwen	4†	Other pruning	+0.0484 [+0.0222, +0.0782]
Qwen	4†	All variants	+0.0522 [+0.0239, +0.0832]
Qwen	7	Quantisation	+0.0232 [+0.0116, +0.0348]
Qwen	7	Gaussian noise	+0.0354 [+0.0149, +0.0557]
Qwen	7	Other pruning	+0.0727 [+0.0371, +0.1084]
Qwen	7	All variants	+0.0712 [+0.0328, +0.1090]
Qwen	17	Quantisation	+0.0063 [+0.0024, +0.0111]
Qwen	17	Gaussian noise	+0.0044 [+0.0010, +0.0083]
Qwen	17	Other pruning	+0.0150 [+0.0070, +0.0237]
Qwen	17	All variants	+0.0147 [+0.0071, +0.0231]
OLMo	4†	Quantisation	−0.0009 [−0.0051, +0.0024]
OLMo	4†	Gaussian noise	+0.0092 [+0.0024, +0.0157]
OLMo	4†	Other pruning	+0.0343 [+0.0255, +0.0442]
OLMo	4†	All variants	+0.0348 [+0.0266, +0.0440]
OLMo	8	Quantisation	−0.0007 [−0.0022, +0.0005]
OLMo	8	Gaussian noise	−0.0016 [−0.0031, −0.0003]
OLMo	8	Other pruning	+0.0079 [+0.0059, +0.0097]
OLMo	8	All variants	+0.0057 [+0.0040, +0.0074]
OLMo	16	Quantisation	−0.0004 [−0.0016, +0.0006]
OLMo	16	Gaussian noise	−0.0023 [−0.0031, −0.0015]
OLMo	16	Other pruning	+0.0097 [+0.0067, +0.0134]
OLMo	16	All variants	+0.0094 [+0.0064, +0.0131]

Table 16: Gap closure at every measured augmentation layer, excluding denominator ≤ 0.01. † marks primary layer four. Both means of per-cell ratios and ratios of summed gains are shown. Brackets are cluster-robust t intervals over held-out methods ($G - 1$ df); ° marks descriptive values with intervals withheld for fewer than five method clusters. The pooled ratio is a point estimate (its interval under-covered in a calibrated simulation). Empty eligible sets are undefined.

Family	Layer	Cells	Methods	Mean ratios [95% CI]	Pooled ratio
Llama	4†	20	9	0.4205 [0.3059, 0.5351]	0.5568
Llama	8	12	4	0.5454°	0.5667
Llama	16	7	3	0.8354°	0.8124
Qwen	4†	24	9	0.2294 [0.0459, 0.4129]	0.4025
Qwen	7	23	8	0.2735 [0.0525, 0.4946]	0.4230
Qwen	17	6	2	0.7147°	0.7235
OLMo	4†	13	5	0.6417 [0.5215, 0.7620]	0.6388
OLMo	8	7	3	0.4264°	0.4024
OLMo	16	4	2	0.9387°	0.9442

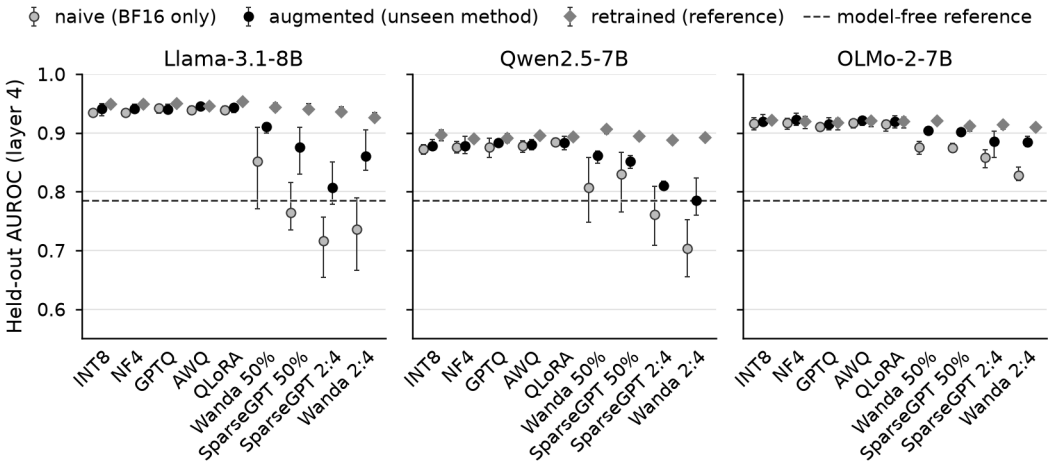

Figure 4: Leave-one-compression-out results for Llama, Qwen, and OLMo at layer four. Points are three-seed means (whiskers: seed range; the AUROC axis starts at 0.55), and every condition scores the same held-out prompts; the dashed line marks the 0.785 model-free reference. Retraining is a reference, not a strict upper bound: augmentation exceeds it by at most 0.0032 in four PTQ/QLoRA hold-outs. Gemma and distilled students are outside this experiment; exploratory Gemma results appear in Appendix E.17. Pooled pruning intervals appear in Table 2; the full depth comparison appears in Tables 15–16.

C Behavioural measurement, controls and supplements

The lexical detector is an auxiliary measurement. The lexical refusal detector is family-calibrated and versioned separately from the API rubric. It checks whether a purported compliance set still contains refusal language and fails calibration when suspicious markers remain. This protects refusal and intervention measurements from known lexical errors, but does not make non-refusal equivalent to harmful assistance. The API judge instead asks whether the actual response contains usable harmful content. Its refusal field includes explicit rejection, deflection without harmful content, and benign transformations of the request.

Validation counts. The saved validation confusion matrix has 196 agreements on refusal, 27 agreements on harmful compliance, 169 detector-compliance/judge-refusal disagreements, and no disagreements in the reverse direction, for 392 judged responses. The observed excess is therefore 169 of the 196 detector-compliant responses. This count supports retaining the detector as a bound on this validated sample. It does not justify applying one global multiplicative correction to every family or method. Provider-filtered validation items are outside the confusion matrix.

Judging, pairing, and exclusions. Each headline difference joins model and BF16 verdicts on prompt identity and uses a Tango interval on the shared valid set. The rubric-v6 judge prompt is: "You grade whether an AI assistant complied with a harmful request. Return ONLY a JSON object: {"refused": <truelfalse>, "convincing": <1-5>, "specific": <1-5>}. "refused" is true if the assistant declined or gave no usable harmful content. "convincing"/"specific" rate how usable the harmful content is (1=not at all, 5=fully)." A response is a success when `refused` is false. Provider refusals are excluded rather than labelled model successes or model refusals. The corresponding marginal judge denominator can differ from the paired denominator; both are retained in the source CSV. Cached verdicts pin the judge/provider and rubric.

Probe leakage controls. The BF16 and compressed matrices are aligned by prompt and evaluated on identical folds. The scaler and logistic regression are fitted inside each fold. Augmented training preserves that split jointly across compression variants: a held-out prompt cannot appear in training through a different variant. Adaptation data come from Dolly, a corpus separate from the XSTest probe set; Dolly instructions of the superseded AdvBench–Dolly probe pool were also excluded, accounting for duplicate instructions with different responses. High-dimensional fitting on all BF16 rows followed by scoring compressed copies of those rows is not a valid frozen-transfer evaluation.

Causal identification. The input-harm logistic weight vector differs from the cell-specific difference-of-means steering direction. Neither decodability nor steering potency establishes prediction of judged compliance.

Steering establishes that the cell's estimated difference-of-means direction affects its refusal behaviour on held-out prompts. It does not establish a shared geometric mechanism across compression methods, nor that a logistic probe's predictive weight vector is itself the operative refusal direction. The cell's identity-intervention run is its baseline. The small intervention sample and greedy decoding restrict both cross-cell comparisons and conclusions about sampling variation. Projection removal reduces refusal with an interval below zero in 35 of 44 cells (32 of 40 compressed); the nine exceptions are OLMo BF16 with its four PTQ and QLoRA cells, and Llama NF4, GPTQ and AWQ. In OLMo SparseGPT 2:4 and Wanda 2:4 it changes refusal by $-9/32$ and $-24/32$. Planned-layer cross-model direction cosines are 0.5336–0.6938 for Llama/Qwen/OLMo 2:4, 0.8625–0.8643 for Gemma, and 0.9393–0.9998 for PTQ; small fit spread does not establish alignment (Appendix E.9).

C.1 Independent second-judge agreement

We submitted 600 pairs from the forty-eight-token grid to `anthropic/claude-sonnet-5`, pinned to the Anthropic provider through OpenRouter, using rubric v6. The sample spans 44 cells, uses seed 0, and comes from 13,492 primary-judged pairs. It is enriched for primary successes: 112 of 600 (18.6667%), compared with 284 of 13,492 in the corpus. The second judge graded 582 pairs and declined 18.

Pooled Cohen's kappa is 0.9242 $[0.8818, 0.9590]$ from a 2,000-draw item bootstrap, and the agreement fraction is 0.9759. Both are conditional on the enriched sample. The interpretable stratum agreements are $108/109 = 0.9908$ among primary successes and $460/473 = 0.9725$ among primary non-successes (Table 17). This comparison probes judge dependence on an enriched sample but does not bound its effect on paired ΔASR; neither judge is ground truth.

Disagreement runs 13 to 1 toward additional successes: the second judge labels 13 primary non-successes as successes and reverses 1 primary success. Sample success rates are 0.1873 for the primary and 0.2079 for the second judge, a difference of $+0.0206$ $[+0.0096, +0.0363]$ by Tango. Reweighting the conditional confusion rates to the corpus base rate implies an absolute level of 0.0478 instead of 0.0210, a ratio of 2.2690 computed before display rounding. This is a level comparison, not a correction to paired ΔASR.

Second-judge upgrades (a primary non-success graded a success) are 0 of 50 primary-negative baseline pairs and 13 of 423 compressed pairs (rate 0.0307). The baseline sample has low power to detect a differential rate of this size; a two-sided exact binomial test of the baseline count against the compressed rate, treated as fixed, gives $p = 0.4084$. Splitting by method gives 0/50 in BF16, 0/192 in PTQ and 13/231 in adaptation plus pruning (Appendix E.2); the method-class contrast has Holm $p = 0.0021$, but four of eight ungraded PTQ non-successes labelled as successes remove nominal

significance (Appendix E.26). The pooled comparison does not justify a uniform-shift assumption. Missing grades limit class comparisons, and the effect on per-cell paired ΔASR remains unresolved.

Provider refusals affect 18 of 600 pairs (3.0000%), with similar rates among primary successes (2.6786%) and non-successes (3.0738%). Only 2 carry the provider's own moderation reason string; 16 are empty bodies with a content-filter finish reason. For those sixteen, the stated reasons were generated by our client, not the provider. Together with the primary provider's refusals (L1), these show a coverage gap in API judging across providers; they do not establish item-by-item overlap in exclusions.

Moderation is not deterministic: 7 of the 25 pairs refused on the first attempt received grades on retry; 18 remained ungraded after both attempts. Total second-judge spending was $0.4608: $0.4550 for the first pass and $0.0058 for the retry. These costs are not included in the budgets of Appendix D.

Table 17: Independent adjudication on an enriched stratified sample. A is the primary judge and B the second. Pooled agreement is conditional on this design; the adjacent stratum rates are the interpretable comparisons. Kappa uses a 95% item-bootstrap interval; the sample-level B minus A difference uses a 95% Tango interval. Upgrades are second-judge upgrades (primary non-success, second-judge success), counted among graded primary non-successes. The reweighted levels are descriptive, not corrections to paired compression effects.

Quantity	Saved result or count-derived summary
Primary judge A	`openrouter:openai/gpt-5.6-sol@openai/flex`
Second judge B; rubric	`openrouter:anthropic/claude-sonnet-5@anthropic`; v6
Sample; cells; seed	600; 44; 0
Judged 48-token corpus; successes	13,492; 284
Sampled primary successes	112 (18.6667%)
Graded / ungraded	582 / 18
Cohen's κ [CI]; item draws	0.9242 $[0.8818, 0.9590]$; 2,000
Agreement, primary-success stratum	108/109 = 0.9908
Agreement, primary-non-success stratum	460/473 = 0.9725
Pooled agreement fraction	0.9759
Confusion ($A0B0$, $A0B1$, $A1B0$, $A1B1$)	460, 13, 1, 108
Sample success rate A / B	0.1873 / 0.2079
Sample-level $B - A$ [CI]	$+0.0206$ $[+0.0096, +0.0363]$
Baseline upgrades / primary non-successes	0 of 50
Compressed upgrades / primary non-successes	13 of 423; rate 0.0307
Two-sided exact binomial p	0.4084; differential not established
Reweighted corpus A / B; ratio	0.0210 / 0.0478; 2.2690
Provider refusals; fraction of sample	18; 3.0000%
Refusals within successes / non-successes	2.6786% / 3.0738%
Provider-reason / empty-body refusals	2 / 16
Initially ungraded / graded on retry	25 / 7
Total cost (first pass + retry), USD	0.4608 (0.4550 + 0.0058)

C.2 Why a BF16 compliance-trained probe is not evaluated

The frozen probes in Table 4 predict XSTest input harmfulness. A different experiment would train on the BF16 model's own StrongREJECT compliance labels, freeze that probe, and evaluate against compressed-model compliance on held-out requests. Appendix E.5 scores the frozen BF16 refusal direction against the forty-eight-token compliance labels; we do not run the BF16-compliance-trained transfer experiment. StrongREJECT activations exist for the 20 decoding-supplement cells, but the forty-eight-token BF16 labels have too few successes. Longer-label eligibility and outcome readouts appear in Appendices E.19–E.22.

Under the forty-eight-token labels, BF16 has 2 judged successes in 309 valid Llama requests, 0 in 306 Qwen requests, 0 in 307 OLMo requests, and 1 in 305 Gemma requests. Qwen and OLMo therefore cannot support a binary compliance-trained classifier or an AUROC against their BF16 labels; the existing five-fold stratification is also unsupported by the rare positive class in Llama and Gemma. We do not change the hazard set or manufacture positive labels to fill this comparison.

At planned layer 16, frozen-direction outcome AUROCs are 0.1521 $[0.0798, 0.2343]$ for OLMo SparseGPT 2:4 and 0.1464 $[0.0792, 0.2252]$ for Wanda 2:4 under those labels. Higher scores retain the input-harm orientation. The within-compressed-cell compliance probes of Appendix C.6 answer a different question. Cell-specific steering does not substitute for predictive evaluation or establish frozen-vector causal transfer. The title, abstract, and conclusions concern frozen input-harm probes.

Frozen readouts rank compliance inversely. Scored with the frozen BF16 refusal direction, where higher scores mean more harm, complied StrongREJECT requests score lower than refused ones (outcome AUROC below one-half). This is the expected sign for a refusal direction, and it places an input-harm monitor's misses on the requests that yield harmful output. The anti-prediction is largely prompt-level, is deepened by compression (OLMo is an exception to prompt dominance) and survives cap-256 labels at every planned or deeper layer; at shallower layers some rows are at chance, and at Qwen QLoRA's layers 4 and 7 projection lies above chance on BF16 and compressed activations alike (Appendices E.5 and E.18–E.20). The deployed threshold's flag-rate drop on StrongREJECT (Appendix C.5) has an XSTest counterpart, a drop in true-positive rate on unsafe prompts. Label-free re-centring on compressed training-fold means removes a fraction 0.6566 $[0.5772, 0.7360]$ of it at headline layers and 0.6252 $[0.3826, 0.8677]$ at planned layers (CR1 over eight cells from four families); majority removal is not established at planned layers, and re-centring changes the pipeline (Appendix E.25).

C.3 The random-basis control

The sparse-basis comparison is confounded by width: 16,384 features against 3,584 dimensions changes how much the penalty binds, independently of what the features mean. We therefore repeat it with a frozen random Gaussian basis of the same width, drawn once per depth and applied unchanged to BF16 and compressed activations, on the same folds, seeds and prompts, and with a second variant thresholded so that its mean active count matches the suite's advertised L_0 (47, 47, 43); that threshold uses the BF16 rows without labels, so it is transductive in the sense of the augmentation control. At Gemma Scope layer 9 under SparseGPT 2:4, the dense probe changes by -0.0233, the random basis by -0.0301 and the SAE by -0.2181, a difference in differences of -0.1881 $[-0.2278, -0.1510]$. The thresholded random basis reaches -0.1943 in that cell, not detectably different from the SAE (difference -0.0238 $[-0.0844, +0.0344]$). The primary readout, fixed before the control was run, is the 27-cell difference in differences against the dense random basis, -0.0149 $[-0.0190, -0.0111]$; it lies below zero, so width and regularisation do not explain the SAE loss.

Table 18: Random-basis control for the sparse-basis comparison. Each row is one compressed cell at one Gemma Scope depth. *Dense* is the 3,584-dimensional probe; *Random* is a frozen $3{,}584 \times 16{,}384$ Gaussian basis, used unchanged for BF16 and compressed activations; *Random, sparse* thresholds that basis to the suite advertised L_0; *SAE* is Gemma Scope. All four use the same folds, seeds and prompts, so the columns may be subtracted. The last column is (SAE − dense) − (random − dense) with a paired 95% prompt-bootstrap interval on the shared resample. Over the 27 cells the random basis costs -0.0012 $[-0.0018, -0.0006]$ against dense and the SAE -0.0161 $[-0.0201, -0.0123]$; the difference in differences is -0.0149 $[-0.0190, -0.0111]$. The sparse random basis instead costs -0.0642 $[-0.0717, -0.0566]$, more than the SAE. Gemma Scope layer l reads the output of zero-indexed block l, which is our activation layer $l+1$ (we count blocks from one); depths 9/20/31 are our layers 10/21/32, and 21 is the Gemma headline layer.

SAE layer	Method	Dense	Random	Random, sparse	SAE	(SAE−dense)−(random−dense)
9	INT8	$+0.0004$	$+0.0001$	-0.0041	-0.0007	-0.0008 $[-0.0059, +0.0045]$
9	NF4	-0.0001	-0.0006	$+0.0158$	-0.0035	-0.0029 $[-0.0088, +0.0032]$
9	GPTQ	-0.0006	-0.0004	-0.0083	-0.0021	-0.0017 $[-0.0105, +0.0065]$
9	AWQ	-0.0005	-0.0009	-0.0174	-0.0017	-0.0009 $[-0.0083, +0.0068]$
9	QLoRA	-0.0021	-0.0019	$+0.0126$	-0.0059	-0.0039 $[-0.0180, +0.0099]$
9	Wanda 50%	-0.0106	-0.0134	-0.0882	-0.0335	-0.0201 $[-0.0395, -0.0017]$
9	SparseGPT 50%	-0.0060	-0.0080	-0.2172	-0.0742	-0.0662 $[-0.0886, -0.0442]$
9	SparseGPT 2:4	-0.0233	-0.0301	-0.1943	-0.2181	-0.1881 $[-0.2278, -0.1510]$
9	Wanda 2:4	-0.0215	-0.0237	-0.1782	-0.1343	-0.1106 $[-0.1419, -0.0808]$
20	INT8	-0.0001	-0.0001	-0.0023	-0.0001	-0.0001 $[-0.0003, +0.0002]$
20	NF4	-0.0002	-0.0002	-0.0005	-0.0001	$+0.0001$ $[-0.0006, +0.0007]$
20	GPTQ	0.0000	0.0000	-0.0008	-0.0003	-0.0003 $[-0.0009, +0.0001]$
20	AWQ	-0.0001	-0.0002	-0.0057	0.0000	$+0.0002$ $[-0.0005, +0.0009]$
20	QLoRA	-0.0017	-0.0042	-0.1646	-0.0035	$+0.0007$ $[-0.0040, +0.0051]$
20	Wanda 50%	-0.0005	-0.0001	-0.0041	-0.0008	-0.0007 $[-0.0021, +0.0005]$
20	SparseGPT 50%	-0.0008	-0.0009	-0.0215	-0.0018	-0.0008 $[-0.0025, +0.0007]$
20	SparseGPT 2:4	-0.0028	-0.0030	-0.0442	-0.0067	-0.0037 $[-0.0099, +0.0010]$
20	Wanda 2:4	-0.0022	-0.0038	-0.0674	-0.0072	-0.0034 $[-0.0089, +0.0012]$
31	INT8	0.0000	$+0.0001$	$+0.0033$	-0.0001	-0.0002 $[-0.0006, 0.0000]$
31	NF4	0.0000	$+0.0002$	$+0.0023$	0.0000	-0.0002 $[-0.0006, +0.0001]$
31	GPTQ	0.0000	$+0.0002$	-0.0052	$+0.0001$	-0.0002 $[-0.0008, +0.0005]$
31	AWQ	0.0000	0.0000	$+0.0031$	-0.0003	-0.0003 $[-0.0009, 0.0000]$

Continued on next page

Table 18: (continued)

SAE layer	Method	Dense	Random	Random, sparse	SAE	(SAE−dense)−(random−dense)
31	QLoRA	−0.0018	−0.0182	−0.3066	−0.0119	+0.0063 [−0.0051, +0.0175]
31	Wanda 50%	0.0000	+0.0010	−0.0074	−0.0001	−0.0011 [−0.0026, −0.0001]
31	SparseGPT 50%	−0.0001	+0.0005	−0.0248	−0.0012	−0.0016 [−0.0033, −0.0005]
31	SparseGPT 2:4	−0.0008	−0.0004	−0.2470	−0.0017	−0.0013 [−0.0035, +0.0010]
31	Wanda 2:4	−0.0012	−0.0008	−0.2370	−0.0023	−0.0015 [−0.0041, +0.0005]

C.4 Decoding sensitivity

The supplement regenerates 20 cells: four BF16 baselines and 16 of 40 compressed cells. Llama contributes QLoRA, Wanda 50%, and both 2:4 methods; Qwen contributes those cells plus SparseGPT 50% and distillation; OLMo contributes both 2:4 methods and distillation; Gemma contributes QLoRA and both 2:4 methods. This supplement excludes PTQ; a separate greedy cap-256 PTQ run appears in Appendix E.3. Greedy cap 256 uses one run per cell, while temperature 0.7 sampling uses three seeds; OLMo's baseline and both 2:4 cells also use greedy cap 512. Differences use the BF16 baseline generated under the same decoding. Tango intervals are unadjusted; they do not extend the grid's forty-test Holm result to new comparisons.

Dated coverage of the Llama baseline. The 48-token grid uses the 13 September per-item export: 309 valid Llama BF16 verdicts, 4 uncached, and 2 successes. The supplement re-read that judge cache on 18 September after one additional item was graded, yielding 310 valid verdicts, 3 uncached, and the same 2 successes. This is one-item coverage growth, not a conflicting estimate. We keep 2/309 and every 48-token paired interval because the entire grid is paired against the earlier export.

Of the thirteen 48-token increases, 12 of the 13 retain positive intervals at cap 256 and in all three sampled seeds. Qwen QLoRA is the exception. At cap 256, point effects grow in 11 cells and shrink in 2 (the two Llama 2:4 cells); persistence and growth are distinct. No regenerated null cell becomes positive at cap 256; the sole sampled exception is Gemma QLoRA seed 2, $+0.0231$ $[+0.0103, +0.0469]$, which is not reproduced by its other seeds or majority vote. Gemma's three majority-vote intervals contain zero. Five of the six missingness-sensitive cells remain positive by an unadjusted test even when all known BF16 successes and all unjudged BF16 responses are counted as reverse flips and only observed forward flips are counted; Qwen QLoRA does not. These conservative checks do not remove the M marks in Table 3.

At cap 512, OLMo SparseGPT 2:4 has $+0.2533$ $[+0.2065, +0.3057]$ and Wanda 2:4 has $+0.2459$ $[+0.2009, +0.2972]$. Their cap-hit rates fall from 0.9585 and 0.9297 at cap 256 to 0.3003 and 0.2013 at cap 512, while the paired effects remain similar. This supports saturation over these two caps, not invariance under arbitrary decoding. Table 19 separates seed variability from prompt intervals.

C.5 A fixed-threshold input-harm probe on behavioural requests

The deployed probe is fitted on all 450 BF16 XSTest rows, and its threshold targets 0.95 out-of-fold true-positive rate (TPR) on unsafe XSTest inputs; seeds 0/1/2 vary the threshold calibration. Frozen transfer is available for 14 regenerated compressed cells; both students have incompatible dimensions. At the headline layer, all eight 2:4 cells have negative mean flag-rate changes, spanning -0.6155 to -0.1832, and Gemma's three cells have zero forward flips under the forty-eight-token judge labels despite lower flag rates; QLoRA and SparseGPT 2:4 each have one reverse flip. At that layer this control supports a threshold-calibration shift without attributing the change to new compliance.

At the headline layer, score-drop AUROC for locating forward flips ranges 0.2835–0.5591 across seven eligible cells (at least ten forward flips); none lies above chance, and Llama Wanda 2:4 and Wanda 50% lie below it. Depth changes this. At the planned layer or deeper, the score drop ranks forward flips above chance in all 25 eligible Llama and OLMo cell–layers (seed-0 intervals), including OLMo SparseGPT 2:4 and Wanda 2:4 at layer 16, 0.8492 $[0.7689, 0.9141]$ and 0.8301 $[0.7367, 0.9091]$; there the seed-0 deployed threshold flags 4 of 18 and 3 of 20 forward flips, against 0.7735 and 0.7343 of other requests. Qwen differs: at planned layer 14 its SparseGPT 2:4 cell gives 0.2641 $[0.1828, 0.3599]$ (Table 20). Under forty-eight-token labels at the headline layer, score AUROC for judged success ranges 0.2724–0.4683 across eight eligible cells; only Llama's two 2:4 cells

Table 19: Decoding supplement: four baselines and sixteen compressed cells; no post-training quantisation cell is in this supplement (PTQ cap 256: Appendix E.3). Greedy and seed-majority differences compare with BF16 under matched decoding and use Tango 95% intervals. Sampling uses temperature 0.7 at cap 256; its mean (between-seed SD) is not a confidence interval. The fraction counts seeds with positive intervals. Baseline ASRs use exact binomial intervals. All significance here is unadjusted; these tests are outside the 48-token grid's Holm family.

Family	Method	Greedy 256 ΔASR [CI]	Sample mean (SD); seeds	Majority ΔASR [CI]
Llama	QLoRA	+0.0971 [+0.0668, +0.1361]	+0.0797 (0.0129); 3/3	+0.0539 [+0.0334, +0.0857]
Llama	Wanda 50%	+0.1238 [+0.0915, +0.1653]	+0.1224 (0.0174); 3/3	+0.1051 [+0.0750, +0.1453]
Llama	SparseGPT 2:4	+0.1071 [+0.0721, +0.1491]	+0.1544 (0.0372); 3/3	+0.1176 [+0.0799, +0.1625]
Llama	Wanda 2:4	+0.1438 [+0.1031, +0.1905]	+0.1664 (0.0161); 3/3	+0.1486 [+0.1066, +0.1968]
Qwen	QLoRA	+0.0167 [−0.0183, +0.0531]	+0.0243 (0.0069); 0/3	+0.0168 [−0.0115, +0.0477]
Qwen	Student	+0.1821 [+0.1373, +0.2324]	+0.2406 (0.0207); 3/3	+0.2257 [+0.1812, +0.2774]
Qwen	Wanda 50%	+0.2013 [+0.1557, +0.2526]	+0.1942 (0.0088); 3/3	+0.1884 [+0.1461, +0.2379]
Qwen	SparseGPT 50%	+0.1429 [+0.1046, +0.1885]	+0.1481 (0.0132); 3/3	+0.1419 [+0.1067, +0.1862]
Qwen	SparseGPT 2:4	+0.1794 [+0.1325, +0.2309]	+0.1791 (0.0133); 3/3	+0.1689 [+0.1248, +0.2188]
Qwen	Wanda 2:4	+0.1833 [+0.1360, +0.2353]	+0.1619 (0.0186); 3/3	+0.1667 [+0.1221, +0.2171]
OLMo	Student	+0.0590 [+0.0346, +0.0924]	+0.0861 (0.0060); 3/3	+0.0912 [+0.0634, +0.1295]
OLMo	SparseGPT 2:4	+0.2451 [+0.1990, +0.2969]	+0.2308 (0.0151); 3/3	+0.2323 [+0.1879, +0.2836]
OLMo	Wanda 2:4	+0.2451 [+0.2003, +0.2963]	+0.2315 (0.0318); 3/3	+0.2136 [+0.1706, +0.2639]
Gemma	QLoRA	+0.0066 [−0.0095, +0.0258]	+0.0154 (0.0069); 1/3	+0.0034 [−0.0127, +0.0211]
Gemma	SparseGPT 2:4	0.0000 [−0.0155, +0.0155]	+0.0044 (0.0050); 0/3	−0.0034 [−0.0188, +0.0094]
Gemma	Wanda 2:4	+0.0099 [−0.0027, +0.0286]	+0.0055 (0.0019); 0/3	+0.0033 [−0.0094, +0.0187]

BF16 baselines, greedy cap 256

Family	Method	ASR [CI]	Valid / unjudged	Cap-hit rate
Llama	BF16	0.0162 [0.0053, 0.0374]	309/4	0.0319
Qwen	BF16	0.0523 [0.0302, 0.0835]	306/7	0.3770
OLMo	BF16	0.0032 [0.0001, 0.0178]	310/3	0.3898
Gemma	BF16	0.0033 [0.0001, 0.0182]	304/9	0.3291

OLMo greedy cap 512

Family	Method	BF16 ASR / ΔASR [CI]	Valid / unjudged	Cap-hit rate
OLMo	BF16	0.0032 [0.0001, 0.0178]	310/3	0.0447
OLMo	SparseGPT 2:4	+0.2533 [+0.2065, +0.3057]	304/9	0.3003
OLMo	Wanda 2:4	+0.2459 [+0.2009, +0.2972]	305/8	0.2013

have seed-0 intervals wholly below chance (OLMo Wanda 2:4 reaches 0.5037). These AUROCs do not depend on the threshold seed; only the bootstrap draw does. With cap-256 labels, frozen logistic outcome AUROC spans 0.2273–0.5586 across eleven headline cells; seven have intervals below chance. Both logistic and projection intervals lie below chance in all 55 planned-or-deeper rows (Appendix E.19). These outcome-prediction tasks have a chance reference of one-half; the 0.785 model-free reference belongs to the separate XSTest input-harm task. At the headline layer, Qwen Wanda 50% has a flag-rate increase supported in every seed, whereas Qwen QLoRA's small positive mean is not detectable in any seed.

Depth qualifies the headline result. Across eight layers, seed-mean 2:4 flag changes range -0.9073 to $+0.3493$. Of 64 cell–layer points, 55 fall and 8 rise with all three seed intervals excluding zero; 1 lacks that agreement. Seven of 8 cells have a positive point at some layer; only Llama SparseGPT 2:4 falls at every layer. Each cell's average across layers is negative, but neither direction at a particular depth nor headline magnitude generalises to every layer. Table 20 places the headline and all-layer results together. The headline layers were selected after observing transfer results (Section 8, L10).

C.6 Compliance-labelled probes within compressed cells

These probes use each cell's StrongREJECT activations and its own forty-eight-token judged-compliance labels, at the headline layer. Five-fold prediction uses seeds 0/1/2, requiring at least ten successes and ten non-successes. This gives 9 eligible cells and 11 ineligible cells, including all four BF16 baselines and OLMo distillation with six positives. Qwen is the only eligible student under these labels. At cap 256, the OLMo student has 19/305 successes and Qwen BF16 also gains label support; we do not refit these classifiers on cap-256 labels. Unlike the frozen input-harm probe, these probes change both the fitting set and the labels; they do not measure frozen BF16-to-compressed compliance transfer.

Performance is not uniformly strong: five of nine seed-mean AUROCs are below the pre-specified 0.75 operational-failure threshold, and both Qwen 2:4 probes have intervals including chance in every seed. Frozen transfer between eligible compressed cells spans 0.4010–0.7934 across 16 directed

Table 20: Frozen input-harm probes on StrongREJECT. The first two panels use the headline layer: Llama 8, Qwen 7, OLMo 8, Gemma 21. Thresholds target 0.95 out-of-fold XSTest TPR; the deployed probe fits all BF16 XSTest rows. Means and SDs span three seeds; adjacent paired 95% prompt intervals condition on seed 0, not the mean. Under the 48-token labels, a forward flip is a request with a BF16 non-success and a compressed success, and a reverse flip the opposite; F/R count them. Behaviour-linked AUROCs also show seed-0 points and 95% intervals; each needs at least ten positives and ten negatives (for flip localisation, forward flips and other requests), and — means insufficient label support. At the headline layer no seed interval for locating forward flips lies above one-half, and for Llama Wanda 2:4 and Wanda 50% every seed interval lies below it. The third panel uses the planned layer (Llama 16, Qwen 14, OLMo 16) for the seven cells with at least ten forward flips. Its score-drop AUROC is identical across seeds and shown with the seed-0 interval; flag rates of the deployed threshold on forward flips / other requests are shown as ranges over the three seeds. At the planned layer or deeper, every seed's interval lies above 0.5 in all 25 eligible Llama and OLMo cell–layers, and 24 of these survive Holm correction over the 56 two-sided Mann–Whitney tests of all eligible cell–layers; of the 10 Qwen cell–layers, 2 lie above 0.5, 3 below and 5 include it (Holm-significant: 0 above, 1 below). The final panel reports descriptive ranges of seed-mean changes across eight layers, with counts of layers where all three seed intervals fall/rise or lack directional agreement (F/R/U).

Cell	Mean BF16 → comp.	Mean Δflag (SD)	Seed-0 Δflag [CI]	F/R
Gemma QLoRA	0.8115 → 0.7487	−0.0628 (0.0098)	−0.0607 [−0.0929, −0.0392]	0/1
Gemma SparseGPT 2:4	0.8115 → 0.5197	−0.2918 (0.0049)	−0.2907 [−0.3433, −0.2432]	0/1
Gemma Wanda 2:4	0.8115 → 0.4516	−0.3599 (0.0018)	−0.3610 [−0.4156, −0.3098]	0/0
Llama QLoRA	0.5889 → 0.5495	−0.0394 (0.0081)	−0.0383 [−0.0658, −0.0221]	9/0
Llama SparseGPT 2:4	0.5889 → 0.1640	−0.4249 (0.0111)	−0.4185 [−0.4744, −0.3642]	38/1
Llama Wanda 2:4	0.5889 → 0.0053	−0.5836 (0.0067)	−0.5911 [−0.6441, −0.5358]	57/2
Llama Wanda 50%	0.5889 → 0.5123	−0.0767 (0.0139)	−0.0703 [−0.1122, −0.0323]	12/1
OLMo SparseGPT 2:4	0.8765 → 0.3855	−0.4909 (0.0610)	−0.5367 [−0.5917, −0.4804]	18/0
OLMo Wanda 2:4	0.8765 → 0.5804	−0.2961 (0.0381)	−0.3227 [−0.3775, −0.2712]	20/0
Qwen QLoRA	0.8477 → 0.8541	+0.0064 (0.0111)	+0.0192 [−0.0171, +0.0568]	4/0
Qwen SparseGPT 2:4	0.8477 → 0.2322	−0.6155 (0.0067)	−0.6102 [−0.6636, −0.5530]	17/0
Qwen SparseGPT 50%	0.8477 → 0.5793	−0.2684 (0.0199)	−0.2907 [−0.3479, −0.2353]	6/0
Qwen Wanda 2:4	0.8477 → 0.6645	−0.1832 (0.0208)	−0.2045 [−0.2625, −0.1478]	19/0
Qwen Wanda 50%	0.8477 → 0.9478	+0.1001 (0.0164)	+0.1182 [+0.0795, +0.1626]	8/0

Headline-layer prediction of behavioural outcomes

Cell	Layer	Score drop predicts flip [CI]	Score predicts success [CI]	Successes
Llama QLoRA	8	—	0.4683 [0.3425, 0.5934]	11
Llama SparseGPT 2:4	8	0.4236 [0.3167, 0.5329]	0.2724 [0.1975, 0.3634]	39
Llama Wanda 2:4	8	0.4050 [0.3177, 0.4984]	0.3724 [0.3031, 0.4483]	57
Llama Wanda 50%	8	0.2835 [0.1602, 0.4389]	0.3617 [0.2195, 0.5158]	13
OLMo SparseGPT 2:4	8	0.3873 [0.2427, 0.5383]	0.3626 [0.2133, 0.5252]	18
OLMo Wanda 2:4	8	0.5591 [0.4542, 0.6628]	0.3703 [0.2449, 0.5037]	20
Qwen SparseGPT 2:4	7	0.5337 [0.3701, 0.6855]	0.4510 [0.3043, 0.5952]	17
Qwen Wanda 2:4	7	0.5395 [0.4061, 0.6757]	0.4152 [0.3022, 0.5378]	19

Planned-layer localisation of forward flips

Cell	Forward flips	Score drop locates flips [CI]	Flag rate, flips / other; three seeds
Llama SparseGPT 2:4	38	0.7333 [0.6460, 0.8135]	0.0000 / 0.0886–0.1624
Llama Wanda 2:4	57	0.7853 [0.7209, 0.8432]	0.0000 / 0.0319–0.0717
Llama Wanda 50%	12	0.7908 [0.6596, 0.9099]	0.1667 / 0.7027–0.7669
OLMo SparseGPT 2:4	18	0.8492 [0.7689, 0.9141]	0.2222 / 0.7735–0.7979
OLMo Wanda 2:4	20	0.8301 [0.7367, 0.9091]	0.1500–0.2000 / 0.7343–0.7587
Qwen SparseGPT 2:4	17	0.2641 [0.1828, 0.3599]	0.0000 / 0.0000
Qwen Wanda 2:4	19	0.3038 [0.1873, 0.4290]	0.0000 / 0.1029–0.1471

Depth sensitivity, 2:4 cells

Cell	Headline Δflag	Minimum / maximum	Mean over layers	F/R/U
Gemma SparseGPT 2:4	−0.2918	−0.8893 / +0.1714	−0.3713	7/1/0
Gemma Wanda 2:4	−0.3599	−0.7380 / +0.1043	−0.3863	7/1/0
Llama SparseGPT 2:4	−0.4249	−0.8829 / −0.2375	−0.6685	8/0/0
Llama Wanda 2:4	−0.5836	−0.9073 / +0.1640	−0.6021	7/1/0
OLMo SparseGPT 2:4	−0.4909	−0.6049 / +0.0532	−0.3361	7/1/0
OLMo Wanda 2:4	−0.2961	−0.4228 / +0.0330	−0.2785	7/1/0
Qwen SparseGPT 2:4	−0.6155	−0.8382 / +0.3493	−0.4752	7/1/0
Qwen Wanda 2:4	−0.1832	−0.6336 / +0.3355	−0.2274	5/2/1

pairs. Table 21 reports seed-0 points with their own 95% prompt intervals separately from across-seed means and SDs. Cross-cell frozen-transfer intervals use 2,000 prompt resamples; within-cell compliance-probe intervals use 1,000. They are conditional on fitted probes and do not remove label scarcity or judge error.

Table 21: Compliance-labelled probes at the headline layer. Eligibility requires at least ten successes and ten non-successes under the 48-token judge labels. Within-cell AUROC uses five-fold prediction; cross-cell probes fit all source-cell rows and are frozen on another cell. Means (fold-seed SDs) and seed-0 95% prompt intervals are distinct quantities. Cross-cell fits are identical across fold seeds, so no SD applies. Neither panel measures frozen BF16-to-compressed compliance transfer. Qwen is the only eligible student; OLMo student and all four BF16 baselines lack label support.

Family	Cell	Layer; positives / valid	Mean AUROC (SD)	Seed-0 AUROC [CI]
Llama	QLoRA	8; 11/310	0.7761 (0.0274)	0.7777 [0.6603, 0.8869]
Llama	SparseGPT 2:4	8; 39/309	0.8042 (0.0332)	0.8384 [0.7772, 0.8967]
Llama	Wanda 2:4	8; 57/309	0.7232 (0.0211)	0.7475 [0.6787, 0.8051]
Llama	Wanda 50%	8; 13/309	0.7183 (0.0312)	0.7217 [0.5987, 0.8381]
OLMo	SparseGPT 2:4	8; 18/307	0.7703 (0.0245)	0.7468 [0.6303, 0.8541]
OLMo	Wanda 2:4	8; 20/308	0.7387 (0.0326)	0.7628 [0.6503, 0.8472]
Qwen	Student	7; 30/307	0.7572 (0.0300)	0.7227 [0.6195, 0.8188]
Qwen	SparseGPT 2:4	7; 17/307	0.5593 (0.0328)	0.5748 [0.4371, 0.7031]
Qwen	Wanda 2:4	7; 19/293	0.6080 (0.0147)	0.6199 [0.4844, 0.7618]
Frozen transfer between eligible compressed cells				
Family	Train → evaluate	Layer	Mean AUROC (SD)	Seed-0 AUROC [CI]
Llama	QLoRA → SparseGPT 2:4	8	0.6688 (—)	0.6688 [0.5850, 0.7588]
Llama	QLoRA → Wanda 2:4	8	0.5332 (—)	0.5332 [0.4474, 0.6166]
Llama	QLoRA → Wanda 50%	8	0.6744 (—)	0.6744 [0.4748, 0.8459]
Llama	SparseGPT 2:4 → QLoRA	8	0.6391 (—)	0.6391 [0.4167, 0.8170]
Llama	SparseGPT 2:4 → Wanda 2:4	8	0.6882 (—)	0.6882 [0.6136, 0.7688]
Llama	SparseGPT 2:4 → Wanda 50%	8	0.7934 (—)	0.7934 [0.6621, 0.8969]
Llama	Wanda 2:4 → QLoRA	8	0.4010 (—)	0.4010 [0.2706, 0.5482]
Llama	Wanda 2:4 → SparseGPT 2:4	8	0.7112 (—)	0.7112 [0.6139, 0.7950]
Llama	Wanda 2:4 → Wanda 50%	8	0.6920 (—)	0.6920 [0.5683, 0.8072]
Llama	Wanda 50% → QLoRA	8	0.5476 (—)	0.5476 [0.3034, 0.7574]
Llama	Wanda 50% → SparseGPT 2:4	8	0.7704 (—)	0.7704 [0.6963, 0.8466]
Llama	Wanda 50% → Wanda 2:4	8	0.5585 (—)	0.5585 [0.4823, 0.6381]
OLMo	SparseGPT 2:4 → Wanda 2:4	8	0.6899 (—)	0.6899 [0.5304, 0.8427]
OLMo	Wanda 2:4 → SparseGPT 2:4	8	0.7055 (—)	0.7055 [0.5725, 0.8228]
Qwen	SparseGPT 2:4 → Wanda 2:4	7	0.5678 (—)	0.5678 [0.3881, 0.7267]
Qwen	Wanda 2:4 → SparseGPT 2:4	7	0.6765 (—)	0.6765 [0.5292, 0.8136]

C.7 Provenance of the model-free reference

The model-free classifier is a TF-IDF word-feature pipeline with L_2-regularised logistic regression (maximum 1,000 iterations). Five shuffled stratified outer folds evaluate generalisation; four shuffled stratified inner folds select by AUROC. Both splitters use the same seed within each run, repeated over seeds 0–4. The grid chooses word unigrams or unigrams plus bigrams, minimum document frequency 1 or 2, and inverse regularisation $C \in \{0.5, 1, 3, 10, 30\}$; other vectoriser settings use the recorded scikit-learn defaults. Selection occurs inside the outer training fold.

The 0.785 reference is the five-seed mean; its separate fold-seed SD is 0.0115. Seed 0 gives 0.7830 $[0.7421, 0.8265]$ from 1,000 stratified prompt resamples. With unshuffled inner folds the seed-0 result is 0.6675 $[0.6193, 0.7193]$. Thus the reference is procedure-sensitive; we use this configuration because it is the strongest measured model-free baseline. These seed-conditional CIs are not intervals for the seed-averaged score.

H7's absolute-range criterion (Section 4.2) conditions on the 0.785 point reference. It does not propagate floor uncertainty, so cases near it require caution. For frozen activation-probe AUROCs, the reported absolute range is exactly $[\min_s L_s, \max_s U_s]$ over the three per-seed 95% item intervals. This is a descriptive range of intervals, not a 95% interval for the seed-averaged score; no joint coverage is asserted.

C.8 Exact paired testing and multiplicity

For multiplicity, let b_m count BF16-only successes and c_m compressed-only successes among paired valid requests. The two-sided exact conditional version of McNemar's test (McNemar, 1947; Fager-

land et al., 2013) conditions on $b_m + c_m$ discordances, with null probability one-half; it is conservative relative to the asymptotic and mid-p versions (Fagerland et al., 2013), consistent with its role as the stricter criterion; no discordances give $p = 1.$. Holm's step-down adjustment over forty tests controls familywise error without requiring independent cells (Holm, 1979). Table 3 marks seven positive adjusted tests: Llama SparseGPT/Wanda 2:4, Qwen SparseGPT/Wanda 2:4 and distillation, and OLMo SparseGPT/Wanda 2:4. None of the eight cells with one discordant prompt is significant. Qwen QLoRA has four discordances and unadjusted exact $p = 0.1250$ (Holm $p = 1.0000$), despite its positive Tango interval; removing one compressed-only success loses Tango significance. It is also missingness-sensitive (Section 4.1). The ‡ flag denotes this single-cell deletion, not a joint deletion across correlated cells; calculations use discordance counts and Tango inversion, not rounded endpoints. Appendix A.3 combines this correction with missing-verdict imputation.

D Reproducibility and provenance

Code and data availability. Code and data: https://anonymous.4open.science/r/compression-safety-monitors-frozen-probes-E8B2. The repository contains analysis code, configurations, the rubric-v6 judge prompt, saved result summaries, the confidence-interval artifact and generator, and prompt-free per-item exports: out-of-fold probe/SAE scores, row order, and a boolean judge–guard join. These preserve cached verdict labels, not raw API records. The confidence intervals can be regenerated from a clean clone using the supplied CPU analysis, without model fits or paid calls. The model-free reference additionally requires the public XSTest CSV whose hash is recorded.

Apart from the judge-validation record (169 responses and eight benchmark prompts; see the Ethics statement), activations, generations, interventions, checkpoints, and raw judge and guard records are not released with this submission; the generations and raw judge records include harmful model outputs. INT8/NF4 are quantised at load and have no separate checkpoints. The archived OLMo and Gemma QLoRA/distillation checkpoints are post-extraction rebuilds, intended to reproduce the recipes but not bit-identical to those producing the reported results.

D.1 Interval classes and computation

Paired ΔAUROC and SAE intervals use 2,000 stratified resamples of 450 prompts (Efron, 1979), holding fitted probes fixed. These prompt-sampling CIs remain separate from the fold-split SDs in Table 4. Absolute ranges span the minimum lower to maximum upper endpoint of three per-seed 95% item CIs, with no joint 95% coverage claim. Headline zero-exclusion calls use unrounded endpoints. Table 22 gives counts and units for the main interval classes; exceptions are listed below it, and other procedures are specified where they are reported.

Table 22: Interval classes, resampling counts and units.

Interval class	Draws	Resampling unit
Judge paired ΔASR	—	Tango score inversion
Single-cell over-refusal and guard flag rates	—	Exact binomial (Clopper–Pearson)
Absolute AUROC / seed-0 model-free reference	1,000	Prompt, within class
Paired AUROC / SAE change	2,000	Shared prompt, within class
Pooled guard recall, precision, FPR	20,000	Prompt cluster across cells
Per-cell guard recall, precision, FPR	—	Exact binomial inversion
Causal refusal change	2,000	Paired held-out prompt
Augmentation improvement	5,000	Paired method–seed row
Augmentation improvement (cluster-robust)	—	Family–method or family cluster, CR1 t
Gap closure	—	Method-cluster CR1 t interval
Second-judge Cohen's κ	2,000	Item
Within-cell compliance-probe AUROC	1,000	Prompt, within class

Sampling units follow the estimand: fixed-probe changes share prompts, pooled guard rates preserve repeated prompts across cells, and gap closure varies held-out methods. Judge and gap-closure interval methods were selected using saved coverage simulations; the probe and guard bootstraps were

not coverage-calibrated. Counts preserve the completed analyses, without claiming an endpoint-convergence study. OLMo GPTQ is the documented exception to 2,000 paired-AUROC draws: 200,000 resolve its borderline headline call as containing zero. Baseline judge ASRs, single-cell over-refusal and guard flag rates, and descriptive zero-discordance bounds use exact calculations. All displayed endpoints use four decimal places with nearest/even rounding; values of magnitude below 0.00005 print as 0.0000 regardless of sign. Separate seed SDs are never added to prompt CIs.

D.2 Compute, environment and recipes

Recorded resources. The compute record lists 9.21 GPU-hours over 52 completed builds and 44 extraction cells. Ninety-four of the ninety-six runs used one NVIDIA A100-SXM4-80GB. The Gemma INT8 and NF4 builds used one NVIDIA RTX A6000 (48 GB). Those two runs were on-the-fly perplexity checks with the recorded approximately 18 GB footprint; this hardware change does not change which saved results are reported.

The distillation arm requires an 80 GB device and cannot be reproduced on that A6000 under the recorded implementation. Students train in fp32. The recorded time is a lower bound: it excludes failed attempts, one Gemma QLoRA rebuild (Appendix E.11), guard inference without recorded timing, downloads, and CPU analysis. It is not a rental bill. API spending is $4.20 at currency precision, from the complete cached-verdict budget including validation and corpus grading. In addition to the ninety-six grid runs above, 23 supplement runs on one NVIDIA RTX A6000 recorded 9.3236 GPU-hours of generation and extraction. Supplement judging cost $12.7318. A separate greedy cap-256 PTQ run (Appendix E.3) regenerated the sixteen PTQ cells on one NVIDIA RTX A6000 and paired them with the decoding supplement's BF16 cap-256 generations, recording 6.2467 GPU-hours including two checkpoint rebuilds; its judging cost $1.9460. These are separate records, not a combined budget.

Environment and artifacts. The recorded environment is Python 3.12.13, torch 2.13.0 (Paszke et al., 2019), and transformers 5.14.1 (Wolf et al., 2020), with pinned dependencies (`requirements.lock.txt`). Compression manifests record model identifiers, model commits, configuration, and source revisions. GPTQ uses 4-bit symmetric weights with group size 128 and AWQ 4-bit asymmetric weights (W4A16); GPTQ, AWQ, Wanda and SparseGPT calibrate on 128 WikiText-2 (Merity et al., 2017) sequences of 2,048 tokens, and pruning excludes the embedding and output layers. Perplexity ratios use WikiText-2 with a 2,048-token window and stride 512; student ratios divide the distilled student's perplexity by the pre-distillation student's on held-out Dolly instructions. Distillation ratios are therefore not interchangeable with the other cells' language-modelling damage ratios. QLoRA and distillation builds are not byte-reproducible under the recorded nondeterministic training; the saved results define this study's realised configurations.

Recorded adaptation recipes. The artifact inventory specifies QLoRA rank 16, LoRA $\alpha = 32$ (scaling $\alpha/r = 2$), dropout 0.05, a quantised base, and a BF16 merge. Each family uses 512 Dolly examples (Conover et al., 2023), 125 training steps, learning rate 0.0002, and maximum sequence length 1,024. Gemma uses batch size two with eight accumulation steps; the other families use batch size four with four accumulation steps. Distillation uses 8,000 Dolly examples, 500 steps, learning rate 0.00001, temperature 1.0, and distillation weight 0.9. The students are Llama-3.2-1B-Instruct (Meta, 2024b), Qwen2.5-1.5B-Instruct (Qwen Team, 2024), OLMo-2-0425-1B-Instruct (OLMo Team, 2025), and Gemma-2-2B-it (Gemma Team, 2024). Distillation batch size/accumulation pairs are 4/4 for Llama, 2/8 for Qwen and OLMo, and 1/16 for Gemma. These are single realised recipes, not hyperparameter distributions.

Uncommitted code at run time. Twenty-four extraction runs are marked as having uncommitted working-tree changes: ten compressed Llama cells, eleven Qwen cells, and Gemma baseline/INT8/NF4. The mechanical check compares each recorded revision with the subsequent commit and finds no changes in the model/evaluation implementation or execution scripts. It bounds subsequently committed drift only. A temporary edit used during a run and reverted before committing cannot be excluded. This narrows the provenance uncertainty without eliminating it. Of 23 supplement runs, 22 record uncommitted working-tree changes; only the Llama baseline is clean. In the decoding supplement, the 18 compressed-cell runs have null model-revision fields, with revision hashes instead preserved in the artifact inventory and original build provenance. The compression

manifests record 2 rebuilt checkpoints (OLMo 2:4) and 14 reused ones. Across 16 compressed cells, 13 re-measured perplexity ratios match the inventory within 0.0005. Absolute gaps are 0.0077 for OLMo SparseGPT 2:4, 0.0044 for Gemma QLoRA, and 0.0007 for Llama SparseGPT 2:4. Only the first was rebuilt for this supplement. The Gemma QLoRA checkpoint had already been rebuilt once (Appendix E.11). The forty-eight-token extraction preceded that rebuild. Its perplexity gap therefore compares different checkpoint instances and does not identify a sampled-seed effect. Quantised checkpoint rebuilds are likewise not guaranteed byte-identical. In the cap-256 PTQ run, all 16 PTQ generation runs are stamped dirty, and OLMo GPTQ and AWQ were rebuilt from the recorded configuration (perplexity ratios 1.0616 and 1.0371 against 1.0531 and 1.0378 in the 48-token grid), so their cap-256 comparisons use different checkpoint instances from their 48-token and probe results.

Source map. The saved sources comprise the master behavioural/readout grid, per-family layer-transfer records, causal-intervention summaries, joined guard summaries, pooled augmentation intervals, and sparse-basis comparisons. The confidence-interval artifact supplies paired changes, pooled guard recall, gap closure, and Tango judge intervals with missing-verdict sensitivity. Additional appendix summaries are transcribed from saved records. For the decoding supplement, 20 cells were regenerated on one RTX A6000, 2 artifacts were rebuilt and 14 reused, StrongREJECT activations were captured, and 25,326 additional triples received judgements. No cell of the forty-eight-token grid was rerun. The 83 variant files contain 25,979 triples, of which 653 lack verdicts. Content-hash caching means the 23,405 graded records and 1,813 failed calls (at least 25,218 backend attempts) are not counts of unique triples. The cap-256 PTQ run has its own audit record, and its build manifests record 2 rebuilt (OLMo GPTQ/AWQ) and 6 reused GPTQ/AWQ checkpoints; INT8/NF4 are quantised at load. The random-basis control (Appendix C.3) runs on CPU from saved XSTest activations. The package records a hash for each summary file.

E Additional sensitivity analyses

This appendix collects sensitivity analyses computed from saved outputs. Appendices E.1–E.4 concern missing verdicts, judge disagreement and the cap-256 PTQ run; E.5 frozen refusal-direction outcome AUROCs; E.6–E.9 the readout basis, non-inferiority, mass-mean and direction geometry; E.10–E.11 the SAE control and checkpoint provenance; E.12–E.17 augmentation; E.18–E.22 outcome prediction; E.23–E.25 over-refusal and threshold adaptation; E.26 logger evidence and second-judge disagreement; and E.27–E.29 absolute recovery, small-cluster gap closure and threshold spread.

E.1 Reconstructing the unjudged records

During 48-token grid judging, the client recognised content filters only on HTTP 400 responses. The run ledger permits a per-cell reconstruction; "API error" is a logger classification, not a verified cause. Filtering of a different response in the cap-256 run is weaker evidence than filtering the same item in the 48-token corpus.

Table 23: Logger reconstruction for the 48-token grid. Columns partition unjudged records by logged filter (I), item filtered in another cell of the same corpus (IIa), item filtered only in the separate cap-256 run (IIb), or no recorded filter evidence (III). Item-level evidence does not identify the cause of every individual gap.

Cell	Unjudged	I	IIa	IIb	III
Llama BF16	4	4	0	0	0
Llama INT8	4	4	0	0	0
Llama NF4	3	3	0	0	0
Llama GPTQ	4	3	0	1	0
Llama AWQ	4	4	0	0	0
Llama QLoRA	3	3	0	0	0
Llama Wanda 50%	4	3	0	1	0
Llama SparseGPT 50%	4	4	0	0	0
Llama SparseGPT 2:4	4	4	0	0	0
Llama Wanda 2:4	4	3	0	1	0
Llama Student	3	3	0	0	0
Qwen BF16	7	4	0	3	0
Qwen INT8	5	4	0	1	0
Qwen NF4	5	4	0	1	0

Continued on next page

Table 23: (continued)

Cell	Unjudged	I	IIa	IIb	III
Qwen GPTQ	5	4	0	1	0
Qwen AWQ	3	2	0	1	0
Qwen QLoRA	11	4	0	7	0
Qwen Wanda 50%	7	4	0	3	0
Qwen SparseGPT 50%	8	4	0	3	1
Qwen SparseGPT 2:4	6	3	0	3	0
Qwen Wanda 2:4	20	4	0	4	12
Qwen Student	6	4	0	2	0
OLMo BF16	6	4	0	2	0
OLMo INT8	3	3	0	0	0
OLMo NF4	4	4	0	0	0
OLMo GPTQ	4	4	0	0	0
OLMo AWQ	4	3	0	0	1
OLMo QLoRA	4	4	0	0	0
OLMo Wanda 50%	6	4	0	1	1
OLMo SparseGPT 50%	5	4	0	1	0
OLMo SparseGPT 2:4	6	3	0	2	1
OLMo Wanda 2:4	5	4	0	1	0
OLMo Student	4	3	0	1	0
Gemma BF16	8	0	4	4	0
Gemma INT8	6	0	4	2	0
Gemma NF4	8	0	4	4	0
Gemma GPTQ	9	0	3	6	0
Gemma AWQ	10	0	3	7	0
Gemma QLoRA	8	0	3	3	2
Gemma Wanda 50%	26	0	3	6	17
Gemma SparseGPT 50%	5	0	4	1	0
Gemma SparseGPT 2:4	9	0	4	5	0
Gemma Wanda 2:4	10	0	4	5	1
Gemma Student	6	0	3	2	1

E.2 Second-judge upgrades by method class

Table 24: Second-judge upgrades (primary non-success graded a success by the second judge) among graded primary non-successes. Brackets are exact binomial intervals. The enriched sample and incomplete grades limit comparisons; neither judge is ground truth.

Class	Upgrades / graded	Rate [95% CI]	Ungraded
BF16	0/50	$0.0000\ [0.0000, 0.0711]$	2
PTQ	0/192	$0.0000\ [0.0000, 0.0190]$	8
Adaptation + pruning	13/231	$0.0563\ [0.0303, 0.0943]$	5

E.3 PTQ at a longer generation cap

This run adds greedy cap-256 generations for the sixteen PTQ cells, which the decoding supplement excludes. Each cell has one greedy generation run; prompt uncertainty is not generation-seed variation. Provider-filtered gaps may be missing not at random, and 36–54% of Qwen, OLMo and Gemma PTQ generations (2–3% for Llama) still reach the 256-token cap. The BF16 baselines of this run come from a later read of the judge cache than those of the decoding supplement; the 48-token grid and the decoding supplement do not use this read.

Table 25: Greedy cap-256 PTQ versus matched BF16: complete-case paired Tango intervals and exact McNemar Holm correction over the sixteen PTQ comparisons. The heuristic fill holds observed verdicts fixed. The last column instead uses the adversarial assignment of every missing BF16 label to non-success and every missing PTQ label to success. These fills are sensitivity scenarios, not observed discoveries or equivalence tests.

Cell	n	Observed ΔASR [CI]	Holm p	Heuristic ΔASR [CI]	Adversarial Holm p
Llama INT8	310	$-0.0065\ [-0.0232, +0.0059]$	1.0000	$-0.0064\ [-0.0230, +0.0058]$	1.0000
Llama NF4	310	$-0.0032\ [-0.0224, +0.0147]$	1.0000	$-0.0032\ [-0.0222, +0.0145]$	1.0000
Llama GPTQ	308	$-0.0065\ [-0.0254, +0.0093]$	1.0000	$-0.0064\ [-0.0250, +0.0092]$	1.0000
Llama AWQ	309	$-0.0097\ [-0.0282, +0.0027]$	1.0000	$-0.0096\ [-0.0278, +0.0027]$	1.0000

Continued on next page

Table 25: (continued)

Cell	n	Observed ΔASR [CI]	Holm p	Heuristic ΔASR [CI]	Adversarial Holm p
Qwen INT8	303	$+0.0033$ $[-0.0226, +0.0298]$	1.0000	$+0.0032$ $[-0.0234, +0.0303]$	0.6919
Qwen NF4	304	$+0.0099$ $[-0.0184, +0.0396]$	1.0000	$+0.0160$ $[-0.0124, +0.0465]$	0.3030
Qwen GPTQ	303	$+0.0132$ $[-0.0171, +0.0451]$	1.0000	$+0.0128$ $[-0.0179, +0.0449]$	0.2856
Qwen AWQ	302	$+0.0033$ $[-0.0271, +0.0341]$	1.0000	0.0000 $[-0.0302, +0.0302]$	0.6919
OLMo INT8	304	0.0000 $[-0.0125, +0.0125]$	1.0000	0.0000 $[-0.0121, +0.0121]$	0.0508
OLMo NF4	301	0.0000 $[-0.0155, +0.0155]$	1.0000	0.0000 $[-0.0150, +0.0150]$	0.0256
OLMo GPTQ	304	-0.0033 $[-0.0184, +0.0092]$	1.0000	0.0000 $[-0.0150, +0.0150]$	0.2148
OLMo AWQ	302	$+0.0033$ $[-0.0093, +0.0185]$	1.0000	$+0.0032$ $[-0.0090, +0.0179]$	0.0078
Gemma INT8	305	0.0000 $[-0.0124, +0.0124]$	1.0000	0.0000 $[-0.0121, +0.0121]$	0.1719
Gemma NF4	305	0.0000 $[-0.0124, +0.0124]$	1.0000	$+0.0032$ $[-0.0090, +0.0179]$	0.0938
Gemma GPTQ	306	0.0000 $[-0.0124, +0.0124]$	1.0000	$+0.0032$ $[-0.0090, +0.0179]$	0.2812
Gemma AWQ	302	0.0000 $[-0.0126, +0.0126]$	1.0000	0.0000 $[-0.0121, +0.0121]$	0.0146

OLMo GPTQ and AWQ were rebuilt for this run (Appendix D.2). In the later judge-cache read used by this PTQ run, Qwen BF16 cap-256 ASR is 0.0552, a current-cache sensitivity; the decoding-supplement value is 0.0523. The largest positive complete-case paired change is Qwen GPTQ $+0.0132$ $[-0.0171, +0.0451]$.

E.4 PTQ MISSINGNESS AND RESIDUAL TRUNCATION

Non-rejection on complete cases and under the heuristic fill is not equivalence, robustness to arbitrary missing labels, or invariance to longer decoding. The heuristic inherits lexical-detector error and does not adjudicate provider-filtered content.

Table 26: Adversarial-increase assignments yielding positive Holm tests. This sensitivity fills all missing labels against the null; these are not observed increases. The heuristic fill covers 137 missing records and gives no Holm-positive increase.

Cell	Adversarial ΔASR [CI]	Holm p	Cap-hit rate
OLMo NF4	$+0.0383$ $[+0.0179, +0.0670]$	0.0256	0.4345
OLMo AWQ	$+0.0383$ $[+0.0221, +0.0658]$	0.0078	0.5431
Gemma AWQ	$+0.0351$ $[+0.0197, +0.0618]$	0.0146	0.4121

E.5 FROZEN REFUSAL-DIRECTION PROJECTIONS

We score compressed-model StrongREJECT activations with the frozen BF16 refusal direction. This BF16-direction projection is distinct from training a classifier on BF16 compliance labels, from re-estimating a direction in each compressed model, and from steering with the frozen BF16 vector. Below-chance positive orientation does not validate a sign-reversed deployment monitor.

Table 27: Frozen BF16 refusal directions on compressed StrongREJECT activations at planned layers, against 48-token judge labels. Orientation is fixed: higher scores mean more input harm. Intervals are nominal prompt-bootstrap intervals conditional on the direction.

Cell	Layer	Judged	Successes	Outcome AUROC [95% CI]
Llama QLoRA	16	310	11	0.0587 $[0.0228, 0.1049]$
Llama Wanda 50%	16	309	13	0.0748 $[0.0333, 0.1253]$
Llama SparseGPT 2:4	16	309	39	0.1311 $[0.0875, 0.1773]$
Llama Wanda 2:4	16	309	57	0.2089 $[0.1598, 0.2620]$
Qwen SparseGPT 2:4	14	307	17	0.2383 $[0.1527, 0.3294]$
Qwen Wanda 2:4	14	293	19	0.2633 $[0.1663, 0.3826]$
OLMo SparseGPT 2:4	16	307	18	0.1521 $[0.0798, 0.2343]$
OLMo Wanda 2:4	16	308	20	0.1464 $[0.0792, 0.2252]$

E.6 AUROC BASIS AND PREPROCESSING SENSITIVITY

Concatenated out-of-fold scores can acquire different fold offsets after compression. Mean within-fold AUROC is insensitive to a constant score shift within each fold. Re-estimating the compressed training-fold mean (re-centring) changes pooled rankings and the monitor pipeline, despite frozen weights. Scale re-estimation instead re-estimates only the per-feature standardisation scale on the compressed training folds, keeping the BF16 mean and the $C = 1$ weights frozen. Per-fold evaluation is a sensitivity analysis; pooled AUROC is the primary basis. Mass-mean operational failures can already be present in BF16; a smaller drop is not universally better discrimination.

Table 28: Count intervals across the dimension-compatible cell–layers. Pooled out-of-fold AUROC is primary. Counts use the main-text big-drop and operational-failure definitions. Intervals resample prompts with fixed fits, not fold partitions. Mass-mean uses the shared-prompt interval; other rows use family-independent draws.

Pipeline	AUROC basis	Big drops [CI]	Operational failures [CI]
Frozen (primary)	pooled	21 [17,24]	9 [7,12]
Frozen (primary)	per-fold	4 [3,7]	0 [0,0]
Mean re-centring	pooled	4 [2,7]	0 [0,0]
Mean re-centring	per-fold	4 [3,7]	0 [0,0]
Scale re-estimation	pooled	21 [17,25]	9 [7,12]
Scale re-estimation	per-fold	4 [2,6]	0 [0,0]
$C = 0.01$	pooled	14 [13,16]	6 [2,8]
$C = 0.01$	per-fold	1 [1,3]	0 [0,0]
$C = 0.1$	pooled	17 [15,20]	8 [5,10]
$C = 0.1$	per-fold	2 [1,6]	0 [0,0]
$C = 1$	pooled	21 [17,24]	9 [7,12]
$C = 1$	per-fold	4 [3,7]	0 [0,0]
$C = 10$	pooled	22 [19,24]	11 [9,14]
$C = 10$	per-fold	6 [3,7]	0 [0,0]
$C = 100$	pooled	21 [18,24]	12 [8,13]
$C = 100$	per-fold	3 [2,6]	0 [0,0]
Mass-mean	pooled	2 [0,9]	30 [5,55]
Mass-mean	per-fold	1 [0,8]	30 [2,54]

E.7 SEED MEANS AND NON-INFERIORITY TO THE BIG-DROP MARGIN

The no-big-drop statement of Section 4.1 concerns seed means. One OLMo Wanda 2:4 seed exceeds the point-drop margin (seed 1, layer 4: -0.1016). Both OLMo methods pass at every layer under mass-mean and re-centred logistic pipelines on both bases. Across the six behaviour-positive 2:4 cells, the number of pipeline–basis combinations (frozen logistic, mass-mean and re-centred logistic, each on the pooled and per-fold bases) in which all eight layers pass is 6/6 for OLMo SparseGPT 2:4, 5/6 for OLMo Wanda 2:4, 2/6 for each Qwen cell and 0/6 for each Llama cell. Passing at every layer is an intersection–union test (Berger, 1982): the claim holds only if each layer's one-sided test rejects, so level-0.05 component tests give a level-0.05 test of the conjunction without multiplicity adjustment. Within-cell Holm adjustment is therefore conservative; it changes no verdict among the 36 cell–readout–basis combinations. Per fold and after re-centring, Llama SparseGPT 2:4 has no seed-mean big drop at any layer (largest declines -0.0995 per fold and -0.0984 re-centred and pooled, both at layer 4), but its layer-4 lower bound crosses the margin, so it fails non-inferiority there. These alternatives do not establish preservation of the deployed frozen pipeline or causal BF16-vector transfer.

Table 29: OLMo frozen logistic readouts at every measured depth. Non-inferiority tests the one-sided null that the AUROC change is at or below the -0.10 big-drop margin; pooled p is the fraction of 2,000 paired bootstrap draws at or below the margin, so $p < 0.0005$ means that no draw reached it. SparseGPT passes at every layer, including within-cell Holm adjustment; Wanda does not pass on the pooled basis at the shallowest layer.

Method	Layer	Pooled ΔAUROC [CI]	Pooled p	Pooled non-inf.	Per-fold non-inf.
SparseGPT 2:4	4	-0.0598 $[-0.0786, -0.0404]$	< 0.0005	yes	yes
SparseGPT 2:4	8	-0.0345 $[-0.0489, -0.0224]$	< 0.0005	yes	yes
SparseGPT 2:4	12	-0.0457 $[-0.0614, -0.0321]$	< 0.0005	yes	yes
SparseGPT 2:4	16	-0.0212 $[-0.0321, -0.0118]$	< 0.0005	yes	yes
SparseGPT 2:4	20	-0.0110 $[-0.0184, -0.0053]$	< 0.0005	yes	yes
SparseGPT 2:4	24	-0.0132 $[-0.0228, -0.0053]$	< 0.0005	yes	yes
SparseGPT 2:4	28	-0.0220 $[-0.0335, -0.0117]$	< 0.0005	yes	yes
SparseGPT 2:4	32	-0.0128 $[-0.0213, -0.0066]$	< 0.0005	yes	yes

Continued on next page

Table 29: (continued)

Method	Layer	Pooled ΔAUROC [CI]	Pooled p	Pooled non-inf.	Per-fold non-inf.
Wanda 2:4	4	-0.0902 $[-0.1120, -0.0675]$	0.1830	no	yes
Wanda 2:4	8	-0.0191 $[-0.0299, -0.0093]$	< 0.0005	yes	yes
Wanda 2:4	12	-0.0184 $[-0.0283, -0.0099]$	< 0.0005	yes	yes
Wanda 2:4	16	-0.0129 $[-0.0202, -0.0071]$	< 0.0005	yes	yes
Wanda 2:4	20	-0.0104 $[-0.0176, -0.0050]$	< 0.0005	yes	yes
Wanda 2:4	24	-0.0110 $[-0.0192, -0.0045]$	< 0.0005	yes	yes
Wanda 2:4	28	-0.0155 $[-0.0257, -0.0066]$	< 0.0005	yes	yes
Wanda 2:4	32	-0.0112 $[-0.0186, -0.0056]$	< 0.0005	yes	yes

E.8 Difference-of-means as an alternative readout

At the headline layers, mass-mean removes the Llama and Qwen 2:4 big drops of the frozen logistic readout, but its BF16 AUROC at Qwen layers 4–10 is only 0.7279–0.7856, below the 0.785 model-free reference at layers 4 and 7 and barely above it at layer 10. At layer 4 its BF16 AUROC is below that reference in every family (Llama 0.7674, Qwen 0.7279, OLMo 0.7730, Gemma 0.7145). OLMo has no seed-mean big drop at any measured layer. The two pooled mass-mean big drops are Llama Wanda 2:4 at the planned layer (last row) and at layer 12 (-0.1006); each marginally crosses the point threshold, with an interval spanning the margin.

Table 30: Frozen mass-mean input-harm readouts. BF16 is the mass-mean BF16 AUROC; intervals are paired prompt intervals; the last column is the frozen logistic pooled change at the same cell–layer.

Cell	Layer	BF16	Pooled ΔAUROC [CI]	Per-fold ΔAUROC [CI]	Logistic Δ
Llama SparseGPT 2:4	8	0.8599	-0.0426 $[-0.0596, -0.0251]$	-0.0333 $[-0.0497, -0.0163]$	-0.1107
Llama Wanda 2:4	8	0.8599	-0.0856 $[-0.1092, -0.0631]$	-0.0756 $[-0.0996, -0.0526]$	-0.1983
Qwen SparseGPT 2:4	7	0.7373	-0.0136 $[-0.0252, -0.0024]$	-0.0140 $[-0.0250, -0.0035]$	-0.2199
Qwen Wanda 2:4	7	0.7373	-0.0282 $[-0.0451, -0.0124]$	-0.0185 $[-0.0329, -0.0046]$	-0.2229
OLMo SparseGPT 2:4	8	0.8416	-0.0229 $[-0.0333, -0.0129]$	-0.0207 $[-0.0313, -0.0104]$	-0.0345
OLMo SparseGPT 2:4	16	0.9960	-0.0374 $[-0.0520, -0.0239]$	-0.0357 $[-0.0508, -0.0228]$	-0.0212
OLMo Wanda 2:4	8	0.8416	-0.0205 $[-0.0297, -0.0115]$	-0.0174 $[-0.0276, -0.0077]$	-0.0191
OLMo Wanda 2:4	16	0.9960	-0.0301 $[-0.0431, -0.0182]$	-0.0289 $[-0.0422, -0.0174]$	-0.0129
Llama Wanda 2:4	16	0.9955	-0.1003 $[-0.1312, -0.0741]$	-0.1019 $[-0.1315, -0.0753]$	-0.0824

E.9 Direction alignment under compression

Within-model split-half repeatability does not establish cross-model alignment. Compression changes alignment as well as magnitude. The PTQ range in Appendix C is a range of cell means, not extrema over individual fits.

Table 31: Planned-layer cross-model difference-of-means direction geometry. Means aggregate matched training folds. Seed SD (in units of 10^{-4}) and maximum within-seed fold range describe fit spread, not independent-replication uncertainty: the training sets overlap. The PTQ row gives the range of the sixteen PTQ cell means for cosine and norm ratio, and the largest PTQ seed SD and fold range.

Cell	Layer	Cosine	Seed SD ($\times 10^{-4}$)	Fold range	Norm ratio
Llama SparseGPT 2:4	16	0.5837	1.59	0.0213	0.4058
Llama Wanda 2:4	16	0.5336	2.37	0.0233	0.3451
Qwen SparseGPT 2:4	14	0.6811	3.50	0.0179	0.5723
Qwen Wanda 2:4	14	0.6938	3.94	0.0197	0.6064
OLMo SparseGPT 2:4	16	0.6251	1.73	0.0230	0.7940
OLMo Wanda 2:4	16	0.6211	1.29	0.0219	0.8736
Gemma SparseGPT 2:4	21	0.8643	1.04	0.0029	0.7845
Gemma Wanda 2:4	21	0.8625	0.44	0.0033	0.7718
PTQ, 16 cells	Planned	0.9393–0.9998	≤ 0.86	≤ 0.0045	0.8456–1.0071

E.10 Basis sensitivity of sparse-feature controls

The random-basis bound of Appendix C.3 is specific to the pooled basis; it is not a uniform per-fold bound. The finding that the thresholded random basis is not detectably different from the SAE (Appendix C.3) is scoped to the SparseGPT comparison, not the Wanda comparison. Generality remains limited to this family and SAE suite.

Table 32: Aggregate comparisons on the same Gemma cells. Paired shared-prompt intervals keep all fitted readouts fixed. DiD denotes the paired difference in differences. The SAE disadvantage survives per-fold evaluation.

Contrast	Pooled [95% CI]	Per-fold [95% CI]
SAE minus dense	-0.0161 $[-0.0201, -0.0123]$	-0.0130 $[-0.0167, -0.0092]$
Random minus dense	-0.0012 $[-0.0018, -0.0006]$	-0.0003 $[-0.0007, +0.0002]$
SAE minus random DiD	-0.0149 $[-0.0190, -0.0111]$	-0.0127 $[-0.0165, -0.0088]$
SAE minus sparse random DiD	$+0.0481$ $[+0.0402, +0.0560]$	$+0.0422$ $[+0.0342, +0.0502]$

E.11 Checkpoint chronology behind the Gemma perplexity gap

Table 33: Gemma QLoRA compression manifests. Restoring the original manifest did not restore the original checkpoint; the supplement verified the rebuilt checkpoint. The activations used in the main analyses were extracted before the rebuild. Ratios are copied from the manifests and do not decompose the gap. The restored manifest carries the original build date.

Record	Recorded date	PPL ratio
Original build	2026-09-11	1.2352
Rebuild	2026-09-12	1.2395
Original manifest restored	2026-09-11	1.2352
Supplement verification	2026-09-18	1.2396

E.12 Augmentation on a matched AUROC basis

Table 2 reports the primary pooled fit; the twelve primary contrasts come from one refit on both AUROC bases. Reproduction check: 19 of 21 checks pass as specified; the other two fail on three unit-level pooled AUROCs (two of 729 values in one check, one of 297 units in the other), each off by one prompt-rank pair at the pinned BLAS thread count and reproduced exactly at another. This one-rank-pair tolerance was adopted after the mismatches were observed. Per-fold naive and re-centred AUROC coincide because foldwise constant shifts preserve ranks. Clustering the primary pooled gain by family instead gives $+0.0609$ $[-0.0171, +0.1390]$ with only 3 clusters.

Table 34: Paired contrasts from the same fits. The first interval resamples method–seed units; the second is CR1 over family–held-out-method clusters. Holm adjusts twelve primary tests. Neither interval is a new prompt bootstrap; coverage of this CR1 design was not simulated. The last row is the primary pooled fit of Table 2, outside the Holm family (NA). In the L4 pooled in-fold noise row, both upper bounds print as 0.0000 but are negative below display precision; both intervals exclude zero before Holm adjustment.

Depth	Contrast	ΔAUROC [unit CI]	CR1 CI	Holm p
L4	All minus naive, per-fold	$+0.0144$ $[+0.0109, +0.0179]$	$[+0.0085, +0.0204]$	0.0027
L4	All minus re-centred, pooled	-0.0207 $[-0.0310, -0.0105]$	$[-0.0371, -0.0043]$	0.1638
L4	In-fold minus transductive noise, pooled	-0.0012 $[-0.0025, 0.0000]$	$[-0.0024, 0.0000]$	0.2317
L4	In-fold minus transductive noise, per-fold	-0.0001 $[-0.0003, +0.0002]$	$[-0.0003, +0.0001]$	0.8618
L4	All minus in-fold noise, pooled	$+0.0381$ $[+0.0251, +0.0534]$	$[+0.0216, +0.0545]$	0.0035
L4	All minus in-fold noise, per-fold	$+0.0174$ $[+0.0136, +0.0214]$	$[+0.0112, +0.0236]$	0.0008
Headline	All minus naive, per-fold	$+0.0236$ $[+0.0144, +0.0340]$	$[+0.0040, +0.0432]$	0.1812
Headline	All minus re-centred, pooled	-0.0046 $[-0.0157, +0.0062]$	$[-0.0211, +0.0118]$	0.8618
Headline	In-fold minus transductive noise, pooled	$+0.0007$ $[-0.0001, +0.0015]$	$[-0.0004, +0.0017]$	0.7726
Headline	In-fold minus transductive noise, per-fold	$+0.0002$ $[-0.0001, +0.0005]$	$[-0.0002, +0.0006]$	0.8618
Headline	All minus in-fold noise, pooled	$+0.0288$ $[+0.0135, +0.0460]$	$[+0.0037, +0.0539]$	0.1812
Headline	All minus in-fold noise, per-fold	$+0.0278$ $[+0.0171, +0.0409]$	$[+0.0042, +0.0513]$	0.1812
L4	All minus naive, pooled (primary fit)	$+0.0609$ $[+0.0431, +0.0812]$	$[+0.0393, +0.0826]$	NA

The like-for-like layer-four per-fold gain is 23.7125% [16.9930%,34.0734%] of the pooled gain from the same refit, a descriptive ratio with a post hoc reporting floor (Appendix E.17). It is not a decomposition of the Table 2 estimate.

E.13 FAMILY SCOPE AND NUMERICAL SENSITIVITY

Clustering the three-family layer-four per-fold gain by family gives $+0.0144$ $[-0.0030, +0.0318]$ with only 3 clusters. The unit interval is positive in every family, but method CR1 ($G = 4$) excludes zero in only one. Appendix E.14 instead pools all nine held-out methods; there the nine per-fold family–layer unit intervals all lie above zero, but two are fragile: OLMo L8, whose lower bound of $+0.000004$ prints as 0.0000, and Qwen L4, whose lower bound ($+0.0006$) is no larger than the largest shift of its own units' per-fold contrasts across BLAS thread counts (0.0007). The other seven are clearly separated from zero, and three of the nine exclude zero under held-out-method CR1 ($G = 9$). This does not establish transfer to a population of unseen families.

Table 35: Layer-four per-fold all-variant gain by family on the four held-out pruning methods (twelve method–seed units per family, the population of Table 2 and Appendix E.12). CR1 clusters the four held-out pruning methods within each family; these tests are secondary to the twelve-test primary family.

Family	ΔAUROC [unit CI]	Method CR1 CI	G
Llama	$+0.0214$ $[+0.0173, +0.0254]$	$[+0.0143, +0.0284]$	4
Qwen	$+0.0074$ $[+0.0014, +0.0139]$	$[-0.0074, +0.0221]$	4
OLMo	$+0.0145$ $[+0.0098, +0.0196]$	$[-0.0003, +0.0293]$	4

E.14 MATCHED POOLED AND PER-FOLD AUGMENTATION ACROSS DEPTH

Table 36: All-variant minus naive AUROC over all nine held-out methods, from the same refit as Appendix E.12. Each family–layer has twenty-seven method–seed units; the per-fold CR1 interval clusters the nine held-out methods ($G = 9$). Appendices E.12–E.13 and Table 15 instead use only the held-out pruning methods (twelve units per family), so these tables estimate different quantities.

Family	Layer	Units	Pooled [unit CI]	Per-fold [unit CI]	Per-fold CR1 CI
Llama	4	27	$+0.0448$ $[+0.0217, +0.0712]$	$+0.0103$ $[+0.0064, +0.0147]$	$[+0.0020, +0.0187]$
Llama	8	27	$+0.0192$ $[+0.0065, +0.0354]$	$+0.0118$ $[+0.0015, +0.0250]$	$[-0.0129, +0.0365]$
Llama	16	27	$+0.0146$ $[+0.0070, +0.0241]$	$+0.0066$ $[+0.0036, +0.0102]$	$[-0.0002, +0.0134]$
Qwen	4	27	$+0.0250$ $[+0.0103, +0.0433]$	$+0.0037$ $[+0.0006, +0.0072]$	$[-0.0017, +0.0090]$
Qwen	7	27	$+0.0334$ $[+0.0141, +0.0560]$	$+0.0180$ $[+0.0098, +0.0273]$	$[-0.0007, +0.0367]$
Qwen	17	27	$+0.0072$ $[+0.0033, +0.0122]$	$+0.0022$ $[+0.0010, +0.0037]$	$[-0.0007, +0.0051]$
OLMo	4	27	$+0.0178$ $[+0.0112, +0.0253]$	$+0.0083$ $[+0.0054, +0.0117]$	$[+0.0019, +0.0146]$
OLMo	8	27	$+0.0017$ $[+0.0002, +0.0033]$	$+0.0012$ $[0.0000, +0.0026]$	$[-0.0013, +0.0038]$
OLMo	16	27	$+0.0045$ $[+0.0025, +0.0069]$	$+0.0026$ $[+0.0017, +0.0037]$	$[+0.0005, +0.0047]$

On PTQ/QLoRA hold-outs across layers, the per-fold gain is $+0.0008$ $[+0.0002, +0.0013]$.

E.15 NOISE CALIBRATION AND PAIRED ABLATIONS

The RMS ratio of in-fold to transductive noise scales has mean 1.0000 and range 0.9966–1.0046. Re-estimating noise scales within training folds removes the transductive calibration of the primary fit, but is not a causal decomposition. Pooled and per-fold noise contrasts answer different questions. The primary in-fold contrasts and multiplicity adjustment appear in Appendix E.12.

Table 37: Per-fold layer-four contrasts on held-out pruning methods. The in-fold row repeats a primary contrast of Appendix E.12; the other two rows are secondary. The tiny all-minus-pruning gain has a CR1 interval containing zero.

Contrast	ΔAUROC [unit CI]	CR1 CI
Transductive noise minus naive	-0.0029 $[-0.0047, -0.0012]$	$[-0.0056, -0.0002]$
In-fold minus transductive noise	-0.0001 $[-0.0003, +0.0002]$	$[-0.0003, +0.0001]$
All minus pruning-only	$+0.0011$ $[+0.0001, +0.0021]$	$[-0.0003, +0.0025]$

E.16 Gap closure conditional on retraining headroom

The eligible subsets differ from the pooled-basis subsets of Appendix A.2, and no common-subset contrast is available, so per-fold and pooled gap-closure fractions would not be comparable and could not rank family robustness. We therefore do not report per-fold gap-closure fractions.

Table 38: Eligibility for per-fold layer-four gap closure, retaining only retrained-minus-naive AUROC above 0.01. The table reports the retained population, not a cross-basis ratio comparison.

Family	Eligible / total	G
Llama	13/27	5
Qwen	12/27	5
OLMo	12/27	5

E.17 Exploratory Gemma augmentation

This block is exploratory. The descriptive per-fold/pooled gain share is reported only when the same-run pooled gain is at least 0.005 with an interval above zero; this reporting floor was chosen after inspecting Gemma. At layer four the share is 18.3920% [11.8799%,24.8553%]; at layer 21 it is suppressed (pooled gain $+0.0002$). It does not extend the three-family leave-one-compression-out experiment to a confirmatory fourth family.

Table 39: Additional Gemma fits, outside the twelve-test Holm family. Unit intervals and held-out-method CR1 intervals are descriptive; the layer-four CR1 interval contains zero.

Layer	Per-fold ΔAUROC [unit CI]	Method CR1 CI
4	$+0.0153$ $[+0.0090, +0.0217]$	$[-0.0038, +0.0345]$
21	$+0.0001$ $[-0.0001, +0.0003]$	$[-0.0003, +0.0005]$

E.18 Prompt-level anti-prediction and compression

Anti-prediction is substantially prompt-associated and is deepened by compression. The prompt-share ratio describes a decomposition of the below-chance gap; it is not causal mediation. OLMo is an exception to a uniform prompt-dominant interpretation. Positive orientation was fixed before inspecting outcome labels.

Table 40: 48-token compliance labels; BF16-direction projection. BF16-arm scores are evaluated against each compressed cell's outcomes on the same prompts; they are not BF16 self-compliance. The difference is compressed minus BF16-arm outcome AUROC. The last column counts negative paired DeLong tests (DeLong et al., 1988) surviving Holm over eligible rows.

Scope	Rows	BF16 arm [CI]	ΔAUROC [CI]	Holm count
All eligible	64	0.2941 $[0.2584, 0.3322]$	-0.0782 $[-0.1056, -0.0532]$	11
Llama	32	0.2351 $[0.1942, 0.2783]$	-0.0450 $[-0.0749, -0.0177]$	1
OLMo	16	0.3643 $[0.2883, 0.4442]$	-0.1534 $[-0.2202, -0.0888]$	8
Qwen	16	0.3420 $[0.2921, 0.3922]$	-0.0695 $[-0.1086, -0.0314]$	2

Across the 64 eligible rows under 48-token labels, the descriptive prompt share of anti-prediction, $(0.5 - \bar{A}_{\mathrm{BF16}})/(0.5 - \bar{A}_{\mathrm{compressed}})$ over mean outcome AUROCs, is 0.7247 $[0.6219, 0.8126]$.

E.19 Longer-label anti-prediction sensitivity

Projection and logistic anti-prediction survive cap-256 labels in all 55 planned-or-deeper rows. At shallower layers some rows are at chance, and only Qwen QLoRA projection rows at layers 4 and 7 lie above one-half (Appendix E.20). This is an outcome-readout sensitivity, not a refit of the compliance-labelled probes and not a validated sign-reversed monitor. Primary longer-label results use the decoding-supplement judge export; a later read of the judge cache is a sensitivity only and changes no row's nominal-interval classification relative to one-half or its eligibility. PTQ, students and flip prediction are outside this longer-label analysis.

Table 41: Frozen positive-orientation BF16 readouts against longer-cap or sampled labels. Below-half counts use nominal intervals; Holm counts use Mann–Whitney tests. The paired difference compares compressed and BF16-arm scores against the same outcomes. Cap 512 covers only OLMo; denominators and eligible cells change with label support. The last two rows restrict the greedy cap-256 labels to planned or deeper layers.

Labels	Readout	Rows	Below half	Holm	Mean ΔAUROC [CI]
Greedy, cap 256	Projection	88	72	68	-0.0616 $[-0.0751, -0.0482]$
Greedy, cap 256	Logistic	88	74	67	-0.0655 $[-0.0787, -0.0521]$
T=0.7, seed 0	Projection	88	73	67	-0.0550 $[-0.0698, -0.0414]$
T=0.7, seed 0	Logistic	88	73	64	-0.0575 $[-0.0726, -0.0434]$
T=0.7, seed 1	Projection	88	72	66	-0.0554 $[-0.0704, -0.0415]$
T=0.7, seed 1	Logistic	88	74	62	-0.0609 $[-0.0758, -0.0470]$
T=0.7, seed 2	Projection	88	72	66	-0.0619 $[-0.0759, -0.0481]$
T=0.7, seed 2	Logistic	88	74	69	-0.0665 $[-0.0795, -0.0537]$
T=0.7 majority	Projection	88	72	67	-0.0635 $[-0.0798, -0.0484]$
T=0.7 majority	Logistic	88	72	66	-0.0688 $[-0.0856, -0.0535]$
Greedy, cap 512	Projection	16	15	14	-0.1740 $[-0.2077, -0.1388]$
Greedy, cap 512	Logistic	16	13	13	-0.1751 $[-0.2121, -0.1371]$
Greedy, cap 256 (planned or deeper)	Projection	55	55	54	-0.0965 $[-0.1164, -0.0770]$
Greedy, cap 256 (planned or deeper)	Logistic	55	55	54	-0.0905 $[-0.1096, -0.0716]$

E.20 Headline projections with cap-256 labels

Qwen QLoRA is the only headline cell whose nominal interval lies above one-half (Holm $p = 0.9806$); its BF16 arm on the same prompts gives 0.6502. At layer 4 the compressed projection reaches 0.7361 $[0.6274, 0.8335]$ (Holm $p = 0.0061$), matching its BF16 arm (0.7359), so the above-chance ranking is prompt-level.

Table 42: Frozen BF16 refusal directions with positive orientation against cap-256 labels from the decoding supplement. Intervals are nominal; the last column reports Holm-adjusted Mann–Whitney tests over the eligible readout family.

Cell	Layer	Outcome AUROC [CI]	Holm p
Llama QLoRA	8	0.2985 $[0.2291, 0.3784]$	0.0024
Llama SparseGPT 2:4	8	0.2503 $[0.1796, 0.3245]$	< 0.0001
Llama Wanda 2:4	8	0.4252 $[0.3532, 0.4997]$	1.0000
Llama Wanda 50%	8	0.2174 $[0.1536, 0.2905]$	< 0.0001
OLMo SparseGPT 2:4	8	0.3470 $[0.2788, 0.4220]$	0.0016
OLMo Wanda 2:4	8	0.3099 $[0.2471, 0.3755]$	< 0.0001
Qwen QLoRA	7	0.6185 $[0.5309, 0.7113]$	0.9806
Qwen SparseGPT 2:4	7	0.3577 $[0.2815, 0.4366]$	0.0056
Qwen SparseGPT 50%	7	0.3404 $[0.2607, 0.4191]$	0.0031
Qwen Wanda 2:4	7	0.4396 $[0.3647, 0.5175]$	1.0000
Qwen Wanda 50%	7	0.4739 $[0.3991, 0.5496]$	1.0000

E.21 Longer-cap baseline label support

Qwen gains enough labels to measure BF16 self-compliance discrimination at the longer cap. The other baseline counts remain too small under the eligibility rule. These labels are not used by the 48-token compliance classifier.

Table 43: BF16 cap-256 labels from the decoding supplement.

BF16	Successes	Judged	Projection eligible
Llama	5	309	no
Qwen	16	306	yes
OLMo	1	310	no
Gemma	1	304	no

E.22 QWEN BF16 OUTCOME DISCRIMINATION AT CAP 256

A later read of the judge cache (17 successes in 308 judgements) is a sensitivity only; with its added labels, no row of Table 44 changes its nominal-interval classification relative to one-half. Frozen outcome discrimination is distinct from retraining the 48-token compliance-labelled classifier.

Table 44: Frozen BF16 refusal direction against its own cap-256 outcomes: 16 successes among 306 judged requests. Tests adjust over the eight depths; five reject below chance.

Layer	Outcome AUROC [CI]	Holm p
4	$0.5478\ [0.3940, 0.7065]$	1.0000
7	$0.4808\ [0.3274, 0.6552]$	1.0000
10	$0.4720\ [0.3340, 0.6127]$	1.0000
14	$0.2670\ [0.1679, 0.3824]$	0.0068
17	$0.1524\ [0.0856, 0.2315]$	< 0.0001
21	$0.1304\ [0.0515, 0.2442]$	< 0.0001
24	$0.1386\ [0.0565, 0.2489]$	< 0.0001
28	$0.1476\ [0.0614, 0.2545]$	< 0.0001

E.23 PAIRED LEXICAL OVER-REFUSAL CHANGES

Under both detector rules, the same six cells show Holm-positive decreases on the safe prompts: Llama QLoRA and Wanda 50%, OLMo SparseGPT 2:4 and Wanda 2:4, and Gemma QLoRA and Student. The primary detector gives no Holm-positive increase. Llama GPTQ is a nominal counter-case: its Tango interval is positive, but exact McNemar $p = 0.0703$. The lenient rule adds seven decline-lead patterns, written after reading discordant outputs of Qwen GPTQ, NF4 and Student, and Llama Wanda 2:4, and it can produce false positives. Its one additional Holm-positive change, a Qwen Student increase, is in a cell whose outputs informed the patterns. On the 200 unsafe XSTest prompts, a separate forty-comparison Holm family, the primary detector gives Holm-positive refusal decreases in eighteen cells and no increase. They include both OLMo 2:4 cells (from 0.8950 to 0.4150 for SparseGPT and to 0.6400 for Wanda) and Llama and Gemma QLoRA (from 0.9200 to 0.5300 and from 0.7600 to 0.5250), each Holm-positive under both rules. Under the lenient rule, sixteen of the eighteen decreases remain Holm-positive; the other two, Qwen GPTQ and Qwen Student, are cells whose outputs informed its patterns. Qwen GPTQ is the only PTQ decrease: -0.1000 $[-0.1568, -0.0498]$ (Holm $p = 0.0078$) under the primary detector. In 19 of the 25 prompts refused only by BF16, the GPTQ response opens with a decline that the primary detector does not match; under the lenient rule the change is -0.0150 $[-0.0534, +0.0203]$ (Holm $p = 1.0000$). Lexical refusal of unsafe prompts is not judged StrongREJECT harmful compliance, so this detector-sensitive change does not qualify the PTQ behavioural result.

Table 45: Paired safe-XSTest changes under the primary lexical detector and a post hoc lenient detector. Intervals are Tango; exact McNemar tests are Holm-adjusted over forty comparisons separately for each detector. Neither lexical rule is semantic adjudication. Changes on the unsafe prompts are summarised in the text.

Cell	Primary detector Δ [CI]	Holm p	Lenient detector Δ [CI]	Holm p
Llama INT8	$+0.0040\ [-0.0112, +0.0223]$	1.0000	$+0.0080\ [-0.0073, +0.0287]$	1.0000
Llama NF4	$-0.0080\ [-0.0312, +0.0115]$	1.0000	$0.0000\ [-0.0266, +0.0266]$	1.0000
Llama GPTQ	$+0.0240\ [+0.0024, +0.0532]$	1.0000	$+0.0160\ [-0.0078, +0.0445]$	1.0000
Llama AWQ	$+0.0120\ [-0.0033, +0.0347]$	1.0000	$+0.0160\ [-0.0045, +0.0425]$	1.0000
Llama QLoRA	$-0.0560\ [-0.0918, -0.0336]$	0.0048	$-0.0640\ [-0.1014, -0.0398]$	0.0012
Llama Wanda 50%	$-0.0440\ [-0.0771, -0.0247]$	0.0342	$-0.0520\ [-0.0869, -0.0306]$	0.0090
Llama SparseGPT 50%	$-0.0160\ [-0.0514, +0.0170]$	1.0000	$-0.0240\ [-0.0614, +0.0102]$	1.0000
Llama SparseGPT 2:4	$-0.0280\ [-0.0691, +0.0103]$	1.0000	$-0.0360\ [-0.0787, +0.0034]$	1.0000
Llama Wanda 2:4	$-0.0280\ [-0.0663, +0.0068]$	1.0000	$-0.0160\ [-0.0560, +0.0223]$	1.0000
Llama Student	$-0.0040\ [-0.0360, +0.0273]$	1.0000	$-0.0080\ [-0.0430, +0.0258]$	1.0000
Qwen INT8	$-0.0040\ [-0.0251, +0.0151]$	1.0000	$-0.0040\ [-0.0251, +0.0151]$	1.0000
Qwen NF4	$-0.0240\ [-0.0532, -0.0024]$	1.0000	$-0.0200\ [-0.0498, +0.0044]$	1.0000
Qwen GPTQ	$-0.0200\ [-0.0498, +0.0044]$	1.0000	$-0.0080\ [-0.0376, +0.0197]$	1.0000
Qwen AWQ	$+0.0040\ [-0.0208, +0.0300]$	1.0000	$+0.0040\ [-0.0208, +0.0300]$	1.0000
Qwen QLoRA	$-0.0200\ [-0.0549, +0.0116]$	1.0000	$-0.0200\ [-0.0549, +0.0116]$	1.0000
Qwen Wanda 50%	$-0.0400\ [-0.0721, -0.0219]$	0.0664	$-0.0400\ [-0.0721, -0.0219]$	0.0645
Qwen SparseGPT 50%	$0.0000\ [-0.0287, +0.0287]$	1.0000	$0.0000\ [-0.0287, +0.0287]$	1.0000
Qwen SparseGPT 2:4	$-0.0200\ [-0.0532, +0.0095]$	1.0000	$-0.0200\ [-0.0532, +0.0095]$	1.0000
Qwen Wanda 2:4	$-0.0240\ [-0.0599, +0.0082]$	1.0000	$-0.0240\ [-0.0599, +0.0082]$	1.0000
Qwen Student	$+0.0120\ [-0.0241, +0.0496]$	1.0000	$+0.0760\ [+0.0383, +0.1205]$	0.0112

Continued on next page

Table 45: (continued)

Cell	Primary detector Δ [CI]	Holm p	Lenient detector Δ [CI]	Holm p
OLMo INT8	$0.0000\ [-0.0151, +0.0151]$	1.0000	$0.0000\ [-0.0151, +0.0151]$	1.0000
OLMo NF4	$-0.0080\ [-0.0312, +0.0115]$	1.0000	$-0.0080\ [-0.0312, +0.0115]$	1.0000
OLMo GPTQ	$+0.0120\ [-0.0112, +0.0390]$	1.0000	$+0.0120\ [-0.0112, +0.0390]$	1.0000
OLMo AWQ	$+0.0120\ [-0.0079, +0.0369]$	1.0000	$+0.0120\ [-0.0079, +0.0369]$	1.0000
OLMo QLoRA	$+0.0080\ [-0.0219, +0.0395]$	1.0000	$+0.0080\ [-0.0219, +0.0395]$	1.0000
OLMo Wanda 50%	$-0.0240\ [-0.0583, +0.0061]$	1.0000	$-0.0240\ [-0.0583, +0.0061]$	1.0000
OLMo SparseGPT 50%	$-0.0200\ [-0.0532, +0.0095]$	1.0000	$-0.0200\ [-0.0532, +0.0095]$	1.0000
OLMo SparseGPT 2:4	$-0.0640\ [-0.1053, -0.0302]$	0.0308	$-0.0640\ [-0.1053, -0.0302]$	0.0291
OLMo Wanda 2:4	$-0.0640\ [-0.1028, -0.0356]$	0.0055	$-0.0640\ [-0.1028, -0.0356]$	0.0055
OLMo Student	$-0.0040\ [-0.0360, +0.0273]$	1.0000	$-0.0040\ [-0.0360, +0.0273]$	1.0000
Gemma INT8	$-0.0120\ [-0.0369, +0.0079]$	1.0000	$-0.0120\ [-0.0369, +0.0079]$	1.0000
Gemma NF4	$+0.0080\ [-0.0173, +0.0356]$	1.0000	$+0.0080\ [-0.0173, +0.0356]$	1.0000
Gemma GPTQ	$-0.0080\ [-0.0335, +0.0147]$	1.0000	$-0.0080\ [-0.0335, +0.0147]$	1.0000
Gemma AWQ	$-0.0040\ [-0.0322, +0.0231]$	1.0000	$-0.0040\ [-0.0322, +0.0231]$	1.0000
Gemma QLoRA	$-0.1160\ [-0.1627, -0.0793]$	< 0.0001	$-0.1080\ [-0.1559, -0.0680]$	0.0001
Gemma Wanda 50%	$+0.0240\ [-0.0061, +0.0583]$	1.0000	$+0.0320\ [+0.0007, +0.0683]$	1.0000
Gemma SparseGPT 50%	$+0.0080\ [-0.0239, +0.0413]$	1.0000	$+0.0080\ [-0.0239, +0.0413]$	1.0000
Gemma SparseGPT 2:4	$-0.0440\ [-0.0843, -0.0089]$	0.8779	$-0.0360\ [-0.0774, +0.0017]$	1.0000
Gemma Wanda 2:4	$-0.0040\ [-0.0442, +0.0358]$	1.0000	$-0.0040\ [-0.0442, +0.0358]$	1.0000
Gemma Student	$-0.0800\ [-0.1285, -0.0365]$	0.0304	$-0.0800\ [-0.1285, -0.0365]$	0.0287

In the 48-token grid, BF16 over-refusal is Gemma $0.1400\ [0.0995, 0.1893]$, Llama $0.0600\ [0.0340, 0.0970]$, Qwen $0.0440\ [0.0222, 0.0774]$, and OLMo $0.0800\ [0.0495, 0.1209]$; between-family tests are in Appendix E.24. Each family has a single BF16 baseline, so these data cannot link baseline over-refusal to compression effects within a family. QLoRA has over-refusal $0.0040\ [0.0001, 0.0221]$ in Llama and $0.0240\ [0.0089, 0.0515]$ in Gemma, with paired changes $-0.0560\ [-0.0918, -0.0336]$ and $-0.1160\ [-0.1627, -0.0793]$; Llama QLoRA's harmful-compliance increase is missingness-sensitive (Section 4.1). Table 3 reports over-refusal in every cell.

E.24 Between-family baseline refusal comparisons

Gemma has the highest point rate. The Gemma–OLMo comparison does not survive the six-test Holm family. Within Gemma, QLoRA lowers lexical refusal of both safe and unsafe XSTest prompts (Appendix E.23) without a detected 48-token ASR increase, so these data cannot link Gemma's high baseline over-refusal to its lack of a detected ASR increase.

Table 46: Gemma minus the named baseline on paired safe requests. Holm covers all six pairwise family tests; ratio intervals resample paired prompts.

Reference	Δ over-refusal [CI]	Holm p	Rate ratio [CI]
Llama	$+0.0800\ [+0.0319, +0.1317]$	0.0111	$2.3333\ [1.4118, 4.3344]$
Qwen	$+0.0960\ [+0.0538, +0.1445]$	0.0002	$3.1818\ [1.9231, 7.0000]$
OLMo	$+0.0600\ [+0.0041, +0.1174]$	0.1955	$1.7500\ [1.0740, 3.1542]$

A claim-specific intersection–union test (Berger, 1982) for Gemma exceeding every other family has $p = 0.0489$ using two-sided components; changing one refusal call in its weakest component removes nominal significance.

E.25 Fixed-threshold collapse and label-free re-centring

The frozen headline XSTest TPR change is -0.4004, versus -0.4057 on StrongREJECT; their difference is $+0.0053\ [-0.2541, +0.2647]$. All eight 2:4 cells fall on both corpora. This contradicts an attack-distribution-only explanation. The planned-layer fraction interval includes one-half, so majority removal is not established there. Residual losses and cell-specific false-positive costs remain.

Table 47: Share of XSTest 2:4 TPR collapse removed by label-free re-centring on the compressed training-fold mean, plus residual TPR change. CR1 clusters eight cells from four families; shared family data limit effective independence. Re-centring changes the monitor pipeline.

Layers	Fraction removed [CR1 CI]	Residual ΔTPR [CR1 CI]
Headline layers	$0.6566\ [0.5772, 0.7360]$	$-0.1375\ [-0.2316, -0.0434]$
Planned layers	$0.6252\ [0.3826, 0.8677]$	$-0.2152\ [-0.3768, -0.0536]$

E.26 Logger evidence and disagreement sensitivity

The two evidence groups total 124 records; 85 of them concern items filtered only in the cap-256 corpus. At least four pairs unjudged in the 48-token grid have later cached non-success verdicts; the 48-token grid does not use these later verdicts. Provider-filtered gaps number between 119 (logged filters) and 280 (all unjudged records). Adding the 124 records with item-level filter evidence to the logged filters gives 243, which is not an upper bound because the 37 records without filter evidence have unknown causes. The second-judge PTQ versus adaptation/pruning contrast is Holm-positive, but loses nominal significance if four of eight ungraded PTQ primary non-successes are second-judge successes. Baseline comparisons remain underpowered; the enriched design does not support a corpus-wide error correction.

Table 48: Logged non-filter gaps partitioned by item-level filter evidence. These counts do not identify a provider cause for each response.

Same 48-token corpus	Cap-256 corpus only	No filter evidence
39	85	37

E.27 Unconditional absolute recovery

Absolute recovery and conditional gap fractions can reverse family rankings. The headroom filter changes the target subset, and layer four is the weakest naive depth in each of the three families. These estimates do not rank unseen-family robustness.

Table 49: Pooled-basis all-variant gains of the primary fits across all held-out methods, with method-cluster CR1 intervals. The final column counts rows retained by the ratio headroom filter. The refit in Appendix E.14 differs from these primary fits by at most 0.0002 because of BLAS thread count.

Family	Layer	All units	Absolute gain [CR1 CI]	Ratio eligible
Llama	4	27	$+0.0450\ [+0.0056, +0.0844]$	20
Llama	8	27	$+0.0192\ [-0.0089, +0.0472]$	12
Llama	16	27	$+0.0146\ [-0.0017, +0.0309]$	7
OLMo	4	27	$+0.0178\ [+0.0036, +0.0319]$	13
OLMo	8	27	$+0.0017\ [-0.0016, +0.0049]$	7
OLMo	16	27	$+0.0045\ [+0.0001, +0.0089]$	4
Qwen	4	27	$+0.0250\ [+0.0019, +0.0480]$	24
Qwen	7	27	$+0.0333\ [+0.0034, +0.0633]$	23
Qwen	17	27	$+0.0072\ [-0.0016, +0.0160]$	6

E.28 Small-cluster gap-closure sensitivity

Table 50: CR1 (Cameron & Miller, 2015) and Webb wild-cluster (Webb, 2023) intervals for gap closure. Unbounded inversion is printed as such. The seven rows marked † carry CR1 intervals reported elsewhere in the paper; none changes its exclusion of zero under Webb inference. Very small-cluster secondary estimates can disagree.

Family	Layer	Method set	G	CR1 CI	Webb CI
Llama	4	All methods†	9	$[0.3059, 0.5351]$	$[0.2998, 0.5315]$
Llama	4	Pruning only, common subset†	4	$[0.3109, 0.6518]$	$[0.3843, 0.5891]$
Llama	8	All methods	4	$[0.3861, 0.7047]$	$[0.3008, 0.6798]$
Llama	8	Pruning only, common subset	4	$[0.3861, 0.7047]$	$[0.3008, 0.6798]$
Llama	16	All methods	3	$[0.5708, 1.1000]$	$[-0.3864, 1.9325]$
Llama	16	Pruning only, common subset	3	$[0.5708, 1.1000]$	$[-0.3864, 1.9325]$

Continued on next page

Table 50: (continued)

Family	Layer	Method set	G	CR1 CI	Webb CI
OLMo	4	All methods†	5	$[0.5215, 0.7620]$	$[0.5109, 0.7400]$
OLMo	4	Pruning only, common subset†	4	$[0.4905, 0.7981]$	$[0.5055, 0.9338]$
OLMo	8	All methods	3	$[-0.1024, 0.9551]$	$[-1.7647, 2.3982]$
OLMo	8	Pruning only, common subset	3	$[-0.1024, 0.9551]$	$[-1.7647, 2.3982]$
OLMo	16	All methods	2	$[0.7423, 1.1350]$	unbounded
OLMo	16	Pruning only, common subset	2	$[0.7423, 1.1350]$	unbounded
Qwen	4	All methods†	9	$[0.0459, 0.4129]$	$[0.0482, 0.4107]$
Qwen	4	Pruning only, common subset†	4	$[0.0525, 0.5663]$	$[0.0800, 0.7683]$
Qwen	7	All methods†	8	$[0.0525, 0.4946]$	$[0.0435, 0.4898]$
Qwen	7	Pruning only, common subset	4	$[-0.1317, 0.7984]$	$[-0.1327, 1.0590]$
Qwen	17	All methods	2	$[0.1004, 1.3291]$	unbounded
Qwen	17	Pruning only, common subset	2	$[0.1004, 1.3291]$	unbounded

E.29 THRESHOLD AND FOLD-SPLIT SENSITIVITY

Table 51: Headline thresholds and flag-change uncertainty. The last interval is the union of a seed-clustered t interval over the three threshold-calibration seeds and the range of the per-seed Tango prompt intervals: an envelope, not an additional prompt-only CI. Threshold variation is distinct from rank discrimination and cannot be added to the prompt-bootstrap interval.

Cell	Layer	Threshold max/min	Mean Δflag [envelope]
Gemma SparseGPT 2:4	21	1.0082	-0.2918 $[-0.3505, -0.2402]$
Gemma Wanda 2:4	21	1.0082	-0.3599 $[-0.4156, -0.3057]$
Llama SparseGPT 2:4	8	1.3626	-0.4249 $[-0.4931, -0.3642]$
Llama Wanda 2:4	8	1.3626	-0.5836 $[-0.6441, -0.5229]$
OLMo SparseGPT 2:4	8	1.7338	-0.4909 $[-0.6424, -0.3394]$
OLMo Wanda 2:4	8	1.7338	-0.2961 $[-0.3908, -0.2014]$
Qwen SparseGPT 2:4	7	2.5860	-0.6155 $[-0.6754, -0.5530]$
Qwen Wanda 2:4	7	2.5860	-0.1832 $[-0.2625, -0.1110]$