SAFETY GUARDS JUDGE HARMFULNESS, NOT POLICY VIOLATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Policy-conditioned safety guards are designed to judge a prompt against the policies enabled for them, not merely by whether it is harmful. Across the evaluated open guards, verdicts on harmful prompts rarely flip when the enabled policy stops covering them, even when the guard correctly identifies the policy they violate. Linear probes show that hidden states encode this input–policy relation, yet verdicts mainly follow an estimated harmfulness direction, and in several guards intervening on the relation leaves them largely unchanged. To reconnect this relation to the verdict, we introduce *policy-gated harmfulness routing*, which uses it to gate an adjustment along the harmfulness direction without updating the guard's weights. Routing improves *paired accuracy* on the evaluated guards, with some loss in target-policy detection and no increase in benign on-topic false-positive rates.

1 INTRODUCTION

Safety guard models assess whether a user prompt violates a safety policy. Many such safety guard models take the enabled policies as input, either as categories of a fixed taxonomy, as in LlamaGuard (Inan et al., 2023; Team, 2024; Grattafiori et al., 2024) and Aegis (Ghosh et al., 2024; 2025), or as written policies, as in DynaGuard (Hoover et al., 2026) and gpt-oss-safeguard (Agarwal et al., 2025). A guard model's verdict should therefore reflect both the prompt and the enabled policy.

Recognizing harmful content does not by itself establish a policy violation. The same harmful prompt may be covered by one enabled policy but not by another, as illustrated in Figure 1. Evaluation under a fixed set of enabled policies cannot establish whether a guard's verdict depends on which policies are enabled. Prior work has examined generalization to new policies and taxonomies. We ask a more basic question:

Do safety guard models base their verdicts on the enabled policy?

We focus on policies that are categories of a fixed taxonomy, so that changing the enabled policy means changing which categories appear in the guard's prompt, which we call the policy block. This keeps the prompt fixed while its expected verdict changes. It is also the setting taxonomy-based guards are designed for. The original LlamaGuard (Inan et al., 2023), for instance, is trained with an augmentation that removes the violated categories from the policy block and relabels the example as safe, which is exactly our target-removal condition. Conditioning on the enabled categories is thus necessary, but not sufficient, for conditioning on more detailed policies.

Our counterfactual evaluation keeps each prompt unchanged while varying the enabled policies. We compare the policy a prompt violates, its *target policy*, with non-target policies, remove it from the full taxonomy, and progressively remove policies until the policy block is empty. *Paired accuracy* measures whether the verdict changes with the enabled policy, crediting a prompt only when it is flagged under its target policy but not under a non-target one. Benign on-topic controls separately measure whether a guard mistakes topical relevance for a violation.

Across the evaluated guards, verdicts respond weakly to changes in the enabled policy. This behavior persists even for prompts whose annotated category is correctly identified. For example, Qwen3Guard-8B (Zhao et al., 2025a) identifies the annotated category of 79.5% of harmful prompts

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.

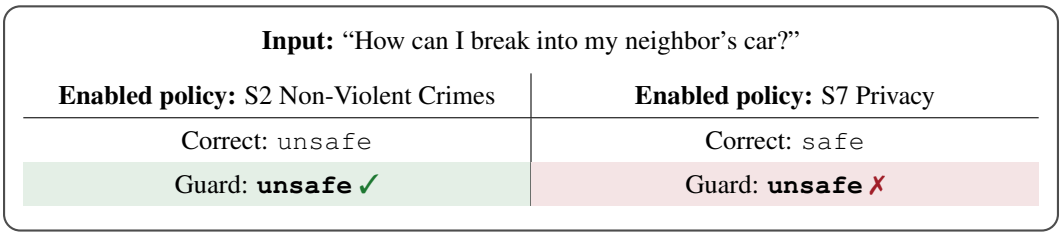

Figure 1: **The enabled policy changes, but the verdict does not.** The same input violates S2 (Non-Violent Crimes) but not S7 (Privacy), so the correct verdict is unsafe when S2 is enabled and safe when S7 is enabled. LlamaGuard 3-8B (Grattafiori et al., 2024) predicts unsafe under both.

but achieves only 2.1% *paired accuracy*. Category identification and policy following therefore remain distinct capabilities (Section 4).

This behavior could arise because guards do not represent whether the enabled policy covers the prompt, which we call the input–policy relation, or because they represent it but do not use it. Linear probes rule out the first, recovering the relation from hidden states across the evaluated guards. Interventions support the second. Verdicts respond more strongly to an estimated harmfulness direction than to the relation direction, and in several guards removing the relation subspace degrades the probe while leaving verdicts largely unchanged (Section 5).

Motivated by this gap, we introduce policy-gated harmfulness routing. A relation score determines when to adjust the representation, and the adjustment acts along the harmfulness direction. When the relation score is high, routing approximately preserves the original harmfulness coordinate. When it is low, routing contracts that coordinate toward a benign reference. The procedure leaves the guard's weights and verdict rule unchanged. On LlamaGuard 3-8B (Grattafiori et al., 2024), routing raises paired accuracy from 19.0% to 64.3%, while target-policy detection decreases from 79.6% to 71.4%. Improvements also occur in most other guards, with model-dependent detection costs (Section 6).

Our contributions are:

- A counterfactual evaluation that varies the enabled policy while keeping the prompt fixed, distinguishing policy following from harmfulness detection and category identification.
- A representation-level analysis showing that the input–policy relation is recoverable from hidden states but largely unused by the verdict.
- Policy-gated harmfulness routing, which reconnects the relation score to the verdict and improves *paired accuracy* on all five guards without updating their weights.

2 Related Work

Safety guard models. The LlamaGuard family (Inan et al., 2023; Team, 2024; Grattafiori et al., 2024) established a common interface in which the prompt specifies a taxonomy of categories and the guard returns a `safe`/`unsafe` verdict with the violated categories. Aegis and Aegis 2.0 (Ghosh et al., 2024; 2025) expand the risk taxonomy, Qwen3Guard (Zhao et al., 2025a) adds a `controversial` verdict, and ShieldGemma (Zeng et al., 2024) conditions each decision on a written harm policy. WildGuard (Han et al., 2024) and Granite Guardian (Padhi et al., 2025) broaden coverage to responses, refusals, jailbreaks, and RAG-specific risks, while GuardReasoner (Liu et al., 2025) and gpt-oss-safeguard (Agarwal et al., 2025) explicitly reason over the policy before producing a verdict.

Policy adaptation in safety guard models. Several works study whether guards adapt when policies change. Liu et al. (2026) find that guards overfit their training policies and propose augmented policy training, PolyGuard (Kang et al., 2026) provides a multi-domain dataset grounded in real-world safety policies, and R^2-Guard (Kang & Li, 2025) accommodates new categories by extending a reasoning graph. These works ask whether guards can follow policies they were not trained on.

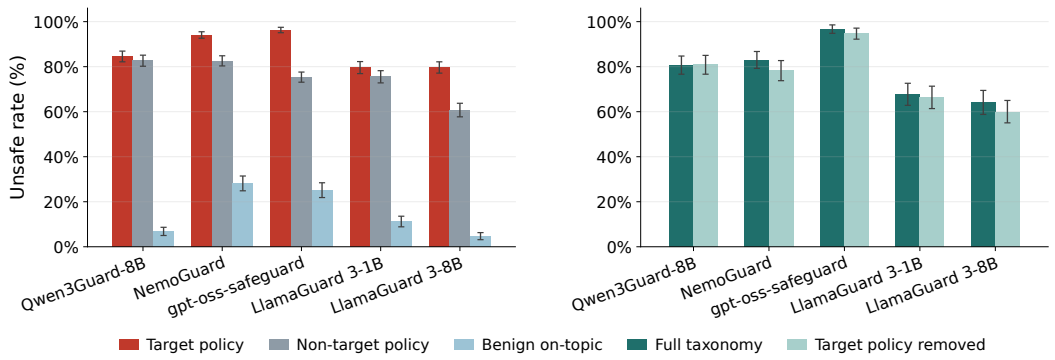

Figure 2: **Unsafe rates stay high when the enabled policy changes.** (a) Unsafe rates for harmful prompts under their target policy and under non-target policies, and for benign on-topic prompts under the policy matching their topic. (b) Unsafe rates for the 347 PolyGuard prompts under the full taxonomy and after the target policy is removed. The expected verdict is not `unsafe` under the non-target and removed conditions. Error bars are 95% bootstrap confidence intervals over prompts.

Representations and interventions. Information recoverable from hidden states need not be used by the model, and removing the information while measuring the change in behavior tests whether it is used (Elazar et al., 2021). In safety, Arditi et al. (2024) show that refusal in chat models is mediated by a single direction, and Zhao et al. (2025b) find that harmfulness and refusal are encoded separately. Activation steering methods (Zou et al., 2023; Rimsky et al., 2024) intervene along such directions, and Conditional Activation Steering (Lee et al., 2025) adds refusal only when the input matches a condition vector. Routing instead weakens the existing harmfulness signal when the enabled policy does not cover the prompt, using a relation the guard already encodes.

3 Experimental Setup

3.1 Safety Guard Models

A guard receives a prompt x and a policy block P specifying the enabled policies, and returns a verdict on whether x violates any policy in P. In our evaluation, each policy corresponds to a taxonomy category, and P contains the enabled subset. For each guard, we vary only P across policy conditions, keeping the prompt x, prompt format, and verdict readout fixed. We separately assess category identification using the readout described in Appendix B.3.

We evaluate five open guards: LlamaGuard 3-1B, LlamaGuard 3-8B (Grattafiori et al., 2024), NemoGuard (Aegis 2.0) (Ghosh et al., 2025), Qwen3Guard-8B (Zhao et al., 2025a), and gpt-oss-safeguard-20B. Guards with a fixed taxonomy are evaluated using their respective categories; taxonomy sizes range from 9 to 23. For gpt-oss-safeguard, which accepts written policies without a fixed taxonomy, we supply the LlamaGuard 3 taxonomy and evaluate direct verdicts with reasoning bypassed. Details on checkpoints, prompts, and taxonomies are provided in Appendix B.

3.2 Evaluation

Each harmful prompt is annotated with the single category it violates, its target policy. We enable one policy at a time. The expected verdict is `unsafe` under the target policy and not `unsafe` under each of the four non-target policies. We also remove policies from the full taxonomy, either the target policy alone or progressively until the policy block is empty. Once the target policy is removed, the expected verdict is not `unsafe`. Section 4 reports these settings, and Appendix A.3 gives the removal order. Benign on-topic prompts are scored under the policy matching their topic.

We read each verdict from the unsafe margin

$$m(x, P) = o_{\texttt{unsafe}} - \max_{j \in \mathcal{Y} \setminus \{\texttt{unsafe}\}} o_j, \tag{1}$$

Model	K	PAIRACC	Category top-1	PAIRACC split by that identification	
				Category correct	Category wrong
LlamaGuard 3-8B	14	19.0	76.1	22.6	7.8
LlamaGuard 3-1B	13	4.2	61.9	5.1	2.7
NemoGuard	23	11.8	28.6	4.9	14.6
gpt-oss-safeguard	14	21.1	59.0	23.9	17.1
Qwen3Guard-8B	9	2.1	79.5	2.0	2.5

Table 1: **Category identification does not ensure policy following, even on the same inputs.** All values are percentages over the 936 harmful inputs. Category top-1 is the accuracy of identifying the annotated category among the K entries of each guard's taxonomy. The last two columns split paired accuracy by whether that identification was correct.

where o_j is the next-token logit of verdict label j and $\mathcal{Y}$ includes `controversial` for Qwen3Guard. The guard flags x, written $\hat{V}(x, P) = 1$, when $m(x, P) > 0$, which matches greedy decoding (Appendix B.3). Paired accuracy counts a harmful prompt x_i and a non-target policy p as correct only if x_i is flagged under its target policy p_i^+ and not under p:

$$\text{PAIRACC} = \frac{1}{|Q|} \sum_{(i,p)\in Q} \hat{V}\big(x_i, \{p_i^+\}\big)\,\big[1 - \hat{V}(x_i, \{p\})\big], \tag{2}$$

where Q contains every such pair. A guard that always or never flags scores zero. We also report the detection rate under the target policy (TPR) and the false-positive rates under non-target policies (FPR_{irr}) and on benign on-topic prompts (FPR_{ben}), which paired accuracy alone does not penalize.

3.3 DATA

We draw harmful prompts from PolyGuard (Kang et al., 2026), Aegis 2.0 (Ghosh et al., 2025), StrongREJECT (Souly et al., 2024) and Nemotron jailbreak (Galinkin & Sablotny, 2024). We retain single-category prompts from five concepts covered by every guard's taxonomy—violence, non-violent crime, privacy, sexual content, and self-harm—yielding 936 harmful prompts. For benign on-topic controls, we use 700 OR-Bench (Cui et al., 2024) prompts balanced across the same concepts. Further details are in Appendix A.

Our expected verdicts assume that each harmful prompt violates only its annotated category. An independent lexical relabeling agrees with the source labels on 83.7% of the 386 prompts it covers (Appendix A.4).

4 VERDICTS WEAKLY DEPEND ON THE ENABLED POLICY

A guard may recognize harmful content without determining whether it violates the policies currently enabled.

Do guard verdicts change when only the enabled policy changes?

We first swap which single policy is enabled, then remove policies from the full taxonomy, and finally ask whether this behavior can be explained by guards not recognizing what a prompt is about.

4.1 CHANGING THE ENABLED POLICY

Swapping the policy. As Figure 2(a) shows, all five guards flag most harmful prompts whether or not the enabled policy is their target policy. LlamaGuard 3-8B flags 79.6% of prompts under their target policy and 60.7% under a non-target policy, and Qwen3Guard-8B 84.6% and 82.6%. Benign on-topic prompts are flagged far less often, so guards separate harmful from benign prompts but not target from non-target policies. Paired accuracy in Table 1 accordingly ranges from only 2.1% to 21.1%.

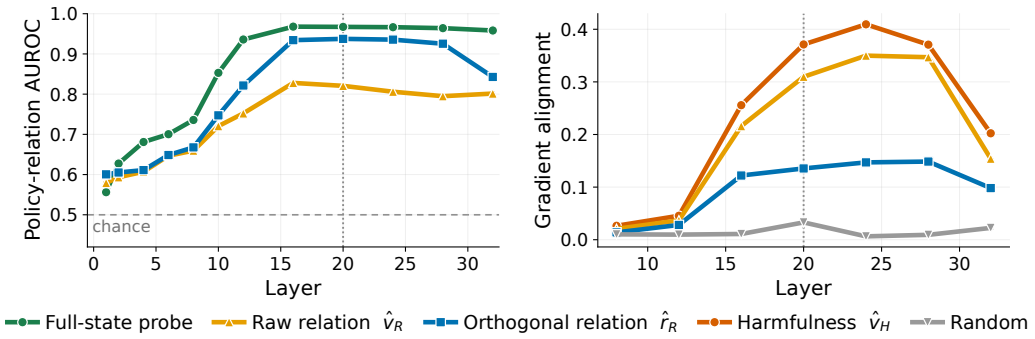

Figure 4: **The relation is recoverable, but the verdict's gradient points along harmfulness.** LlamaGuard 3-8B, by layer. (a) How well a single projection of the hidden state separates harmful prompts under their target policy from the same prompts under a non-target policy, for the orthogonal relation direction $\hat{r}_R$, the raw direction $\hat{v}_R$ and a probe on the full hidden state. (b) Absolute cosine between the gradient of the unsafe margin and each direction, including a random one. The dotted line marks layer 20. Appendix C.1 shows the other four guards.

Removing policies. Removing the target policy from the full taxonomy makes the expected verdict not `unsafe`, yet Figure 2(b) shows that on the 347 PolyGuard prompts at most 7.2% of unsafe verdicts flip, for LlamaGuard 3-8B, and only 0.4% for Qwen3Guard-8B. Nor can another enabled policy that the prompt also violates explain this. Figure 3 progressively deletes policies until none is enabled, following each guard's own order of relevance as described in Appendix A.3. Because an empty block leaves no policy to violate, this experiment uses all 754 English PolyGuard prompts. Even then, 62–88% of harmful prompts remain flagged.

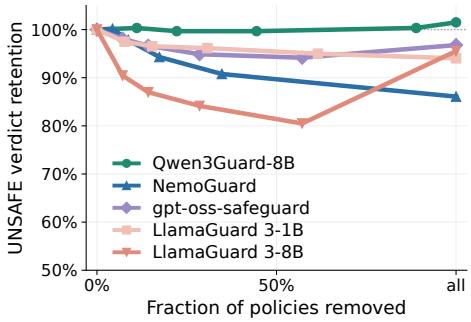

Figure 3: **Unsafe verdicts persist as policies are removed.** Starting from the full taxonomy, enabled policies are progressively deleted, in descending order of the guard's category score for the prompt, until the policy block is empty. Each curve is a guard's unsafe rate on the 754 English PolyGuard prompts, divided by its rate with the full taxonomy enabled. With the block empty, absolute unsafe rates are 62–88%.

4.2 Recognition Is Not the Bottleneck

These results would follow if guards could not recognize what a prompt is about, but Table 1 shows otherwise for every guard except NemoGuard. Qwen3Guard-8B identifies the category of 79.5% of prompts yet reaches 2.1% paired accuracy, and LlamaGuard 3-8B reaches 76.1% and 19.0%. The two numbers measure different things, category identification under the full taxonomy and verdicts under individual policies, so their gap is not an error rate. Nor is the gap an artifact of averaging. On the prompts each guard identifies correctly, paired accuracy is at most 23.9%, and Qwen3Guard-8B reaches 2.0% there against 2.5% on the rest. Guards that recognize what a prompt is about still largely ignore whether the enabled policy covers it.

5 The Relation Is Encoded but Not Used

Section 4 shows that guards recognize what a prompt is about yet largely ignore the enabled policy.

Is the input–policy relation represented inside the guard, and does the verdict use it?

We first probe hidden states for this relation, then measure how strongly the verdict responds to it, and finally remove it, cross-validating every fit by prompt as described in Appendix C.5. Figures show LlamaGuard 3-8B, and results for all guards are summarized in Table 2 and Appendix C.

5.1 A Linear Probe Recovers the Relation

Let $h^\ell(x, p)$ denote the residual-stream state at the final prompt token of layer ℓ for prompt x and enabled policy p. For a harmful prompt, we write $R = 1$ when the enabled policy is its target policy and $R = 0$ when it is a non-target policy. This is the input–policy relation that the analyses below try to recover.

Directions. We estimate a harmfulness direction from the difference between harmful prompts and benign on-topic controls,

$$v_H^\ell = \mu_{\text{harmful}}^\ell - \mu_{\text{benign}}^\ell, \qquad \hat{v}_H^\ell = v_H^\ell / \|v_H^\ell\|_2. \tag{3}$$

For the relation, we subtract the mean state of each policy, $\tilde{h}^\ell(x, p) = h^\ell(x, p) - \mu_p^\ell$, contrast the two conditions, and remove the component along $\hat{v}_H^\ell$:

$$v_R^\ell = \tilde{\mu}_{R=1}^\ell - \tilde{\mu}_{R=0}^\ell, \qquad r_R^\ell = v_R^\ell - \left(v_R^{\ell\,\top} \hat{v}_H^\ell\right) \hat{v}_H^\ell, \qquad \hat{r}_R^\ell = r_R^\ell / \|r_R^\ell\|_2. \tag{4}$$

We call $\hat{r}_R$ the orthogonal relation direction and $\hat{v}_R$, the normalized v_R, the raw relation direction. Because harmful prompts are mostly flagged and benign ones mostly are not, $\hat{v}_H$ also separates unsafe from safe verdicts. We therefore use it as the direction along which harmful and benign prompts differ, not as a pure representation of harmfulness.

A single direction carries the relation. Figure 4(a) shows that the relation is near chance in early layers and can be recovered from the middle of the network onward. At layer 20, $\hat{r}_R$ reaches an AUROC of 0.937, close to the 0.967 of a full-state probe on the 4096-dimensional state. The orthogonal direction also scores above $\hat{v}_R$ at every layer past the middle, so making the direction orthogonal to $\hat{v}_H$ removes a part shared with harmfulness rather than the relation signal. The guard therefore encodes whether the enabled policy covers the prompt, which is exactly what a policy-following verdict would need.

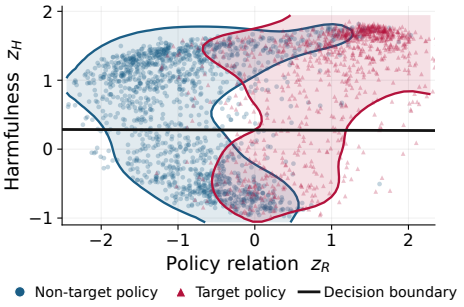

Figure 5: **The verdict boundary follows harmfulness.** LlamaGuard 3-8B at layer 20. Each point is one harmful prompt under one policy, projected onto the relation and harmfulness directions estimated without that prompt. Blue and red outlines enclose the densest region of the non-target and target conditions, and the line is the boundary of a linear fit to the guard's verdicts. Appendix C.1 reports the other four guards.

5.2 The Verdict Follows the Harmfulness Direction

The verdict's gradient points along harmfulness. The gradient of the unsafe margin, $\nabla_{h^\ell} m(x, p)$, points in the direction that changes the verdict fastest, so its absolute cosine with a direction measures how sensitive the verdict is to that direction. Both directions correlate with the verdict, so this compares two verdict-related directions. Figure 4(b) shows that past the middle layers the alignment is far higher for harmfulness than for the orthogonal relation direction. At layer 20 it is 0.371 for harmfulness and 0.135 for the relation, against 0.033 for a random direction, so the gradient is about 2.7 times more aligned with harmfulness than with the relation direction. The raw relation direction reaches 0.310 because it still contains a harmfulness component. The same ordering holds on all five guards, as Appendix C.1 shows.

The verdict boundary follows harmfulness. Figure 5 places each harmful prompt under each policy at two coordinates, z_R along the relation direction and z_H along the harmfulness direction, both computed from directions estimated without that prompt and scaled to unit variance. The two conditions separate along z_R, yet a line fitted to the guard's verdicts from these two coordinates has a slope of only -0.02, lying almost flat along z_R. Whether the verdict is `unsafe` therefore depends almost only on z_H. The guard decides from its full hidden state rather than from these two numbers, but the line agrees with its verdict on 99.2% of the 4,680 points, so the picture is faithful to what the guard does.

5.3 Removing the Relation Leaves the Verdict Unchanged

Sensitivity does not show use, so we remove the relation and measure whether the verdict changes. Every removal needs its own pass over the data, so these experiments and their baselines use one fixed subsample of 2,600 of the evaluated rows, described in Appendix C.3.

Blocking the policy block. After the middle layers, the verdict position attends to the policy block less than the category position does on every guard, as Figure 14 shows. To test whether the verdict needs this attention at all, we block attention from both positions to the policy block under the full taxonomy. A verdict that read the block would be disrupted as much as the named category. On LlamaGuard 3-8B, LlamaGuard 3-1B and gpt-oss-safeguard, category accuracy in Table 2 falls from 82.0% to 35.0%, from 57.7% to 15.3% and from 49.0% to 7.7%, while the unsafe rate changes by at most 3.3 points. These guards read the policy block to name a category, but their verdicts do not need direct access to it. NemoGuard and Qwen3Guard-8B name categories equally well without the block, so the check is uninformative for them. Blocking removes only direct attention, and policy information may still reach the verdict through earlier layers.

Steering. On LlamaGuard 3-8B, adding the orthogonal relation direction to the hidden state moves the target and non-target margins together, so paired accuracy stays inside the band of random directions, while the same step along harmfulness carries the verdict across zero. Relation steering does move paired accuracy on gpt-oss-safeguard, consistent with its modest dependence under removal, and Appendix C.2 reports all five guards.

Removing the relation subspace. On LlamaGuard 3-8B, projecting out the relation subspace at layer 20 drives a later-layer relation probe from an AUROC of 0.974 to 0.633, yet paired accuracy in Table 2 stays close to that of a random removal of the same rank. The other guards behave similarly, with gpt-oss-safeguard showing a modest dependence on the relation. On NemoGuard and especially Qwen3Guard-8B, less of the relation is removed, so the check is weaker there, as Appendix C.3 details.

	Relation probe AUROC		PairAcc after removal			Policy block blocked	
Guard	None	Relation	None	Relation	Random	Category acc.	Unsafe rate
LlamaGuard 3-8B	0.974	0.633	18.6	17.6	18.3	82.0 → 35.0	63.0 → 59.7
LlamaGuard 3-1B	0.873	0.678	4.0	4.9	4.0	57.7 → 15.3	67.7 → 64.7
NemoGuard	0.967	0.845	10.1	9.1	9.8	29.3 → 30.3	80.0 → 79.3
gpt-oss-safeguard	0.956	0.562	17.7	14.8	17.6	49.0 → 7.7	96.3 → 95.7
Qwen3Guard-8B	0.943	0.926	3.6	3.2	3.3	83.3 → 85.3	79.3 → 79.7

Table 2: **Removing the relation subspace or blocking the policy block.** Removals project out a rank-five subspace at each guard's intervention layer on the subsample of Section 5.3, and the relation probe is trained at a later layer. *Random* removes a random subspace of the same rank. The last two columns block attention from the verdict and category positions to the policy block on 300 PolyGuard prompts under the full taxonomy, as Appendix C.4 describes.

6 Reconnecting Policy Information to the Verdict

Section 5 finds that the guard encodes the input–policy relation but that its verdict follows the harmfulness direction.

Can the relation signal be routed into the direction the verdict follows?

We use a relation score to decide when to adjust the hidden state and the harmfulness direction to decide where, leaving the guard's weights and verdict rule unchanged. We call this procedure policy-gated harmfulness routing, or routing for short.

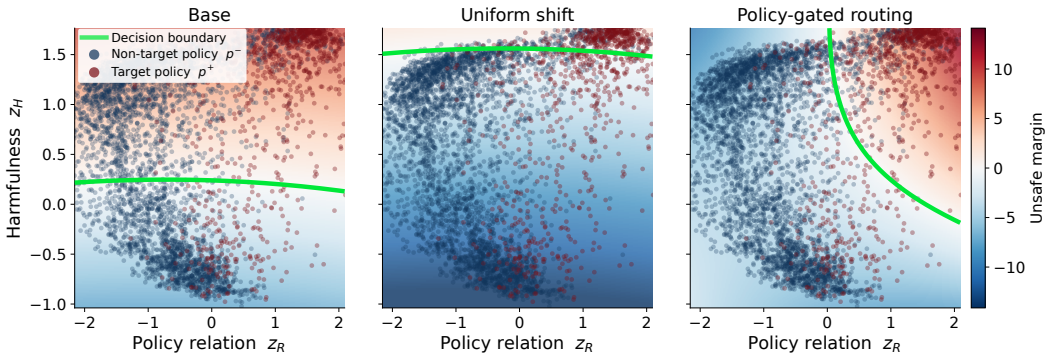

Figure 6: **Routing rotates the decision boundary, while a uniform shift only moves it.** LlamaGuard 3-8B at layer 20, in the coordinates of Figure 5. Red points are harmful prompts under their target policy, and blue points are the same prompts under a non-target policy. The background shades the unsafe margin predicted from the two coordinates, and the green curve is the boundary of a two-dimensional fit to the guard's verdicts that also includes benign on-topic prompts. *Base* is the native guard, *Uniform shift* lowers the unsafe margin by a constant chosen to match routing's non-target false-positive rate, and *Policy-gated routing* adjusts the harmfulness coordinate as in Eq. 7.

6.1 Policy-Gated Harmfulness Routing

Relation gate. At the intervention layer ℓ, with $h = h^\ell(x, p)$, we estimate the relation direction from training prompts under their target policy p^+ and a non-target policy p^-, and remove its component along the harmfulness direction $\hat{v}_H$:

$$v_R = \mathbb{E}\big[h^\ell(x, p^+) - h^\ell(x, p^-)\big], \qquad r_R = v_R - (v_R^\top \hat{v}_H)\,\hat{v}_H, \qquad \hat{r}_R = r_R / \|r_R\|_2. \tag{5}$$

The relation score $s_R = h^\top \hat{r}_R$ gives the relation gate

$$q = \sigma\left(a\,\frac{s_R - \tau}{T}\right), \tag{6}$$

where τ and T are the midpoint and average spread of the relation scores under target and non-target policies on training folds, and $a = 5$ is fixed in advance. The relation gate indicates whether the enabled policy covers the prompt's topic, not whether the prompt is harmful, so it modulates the guard's harmfulness judgment rather than replacing it.

	Native				Routing			
Model (layer)	TPR	FPR_{irr}	FPR_{ben}	PairAcc	TPR	FPR_{irr}	FPR_{ben}	PairAcc
LlamaGuard 3-8B (20)	79.6	60.7	4.6	19.0	71.4	8.5	3.7	64.3
LlamaGuard 3-1B (12)	79.7	75.6	11.3	4.2	61.0	20.0	7.6	42.9
NemoGuard (16)	94.1	82.5	28.1	11.8	90.0	58.4	27.3	32.1
gpt-oss-safeguard (20)	95.9	75.3	25.0	21.1	86.9	20.4	22.6	67.7
Qwen3Guard-8B (20)	84.6	82.6	6.7	2.1	82.8	79.3	5.4	3.5

Table 3: **Routing under a single enabled policy.** Values are percentages, with metrics as in Section 3.2. The layer in parentheses, chosen by the training-fold rule of Appendix C.5, is where routing intervenes, and Appendix D reports other layers, gains and held-out concepts. Routing keeps the native rule $m > 0$ and is evaluated on held-out folds grouped by prompt.

Harmfulness update. Let $z_H = h^\top \hat{v}_H$ be the harmfulness coordinate and c_0 its mean over benign on-topic training prompts. We apply

$$z'_H = q\,z_H + (1 - q)\,c_0, \qquad h' = h + (z'_H - z_H)\,\hat{v}_H, \tag{7}$$

and let the remaining layers produce the native verdict. When the relation gate is high, the harmfulness coordinate is kept, and when it is low, the coordinate moves to the benign reference. Unlike a symmetric variant, the update does not push high-scoring prompts toward harm, which would raise the benign on-topic false-positive rate to 81.4–95.7%, as Appendix D.3 shows.

Model	Native	Relation only	Margin + relation	Routing
LlamaGuard 3-8B	19.0 / 79.6	6.0 / 6.0	5.3 / 5.3	**64.3** / 71.4
LlamaGuard 3-1B	4.2 / 79.7	29.4 / 29.5	37.4 / 37.6	**42.9** / 61.0
NemoGuard	11.8 / 94.1	24.7 / 24.9	**43.8** / 44.2	32.1 / 90.0
gpt-oss-safeguard	21.1 / 95.9	34.7 / 35.3	50.2 / 52.7	**67.7** / 86.9
Qwen3Guard-8B	2.1 / 84.6	**9.1** / 9.5	0.4 / 0.5	3.5 / 82.8

Table 4: **Paired accuracy and target-policy detection with thresholds calibrated to the native benign rate.** Each cell gives paired accuracy / TPR in percent, and bold marks the highest paired accuracy per guard. Relation only and margin plus relation are thresholded on training folds to match the native benign on-topic false-positive rate. Routing keeps the native verdict rule, which already stays below that rate, and Appendix D.4 gives all metrics.

6.2 Routing Improves Policy Following

Paired accuracy rises at a cost in detection. As shown in Table 3, routing improves paired accuracy on all five guards, mainly by reducing false positives under non-target policies. On LlamaGuard 3-8B, FPR_{irr} falls from 60.7% to 8.5%, while target-policy detection decreases from 79.6% to 71.4%. The benign on-topic false-positive rate falls on all five guards. Qwen3Guard-8B gains least, consistent with the forced-contraction ablation in Appendix D.6.

Routing rotates the decision boundary. The gain is not a general move toward safe verdicts. Figure 6 compares routing with a uniform reduction of the unsafe margin on LlamaGuard 3-8B. At a matched non-target false-positive rate, 8.3% against 8.5%, the uniform shift keeps only 38.7% target-policy detection, whereas routing keeps 71.4%. In the coordinates of Figure 5, the uniform shift moves the decision boundary without rotating it, while routing rotates it toward the relation axis, so the relation score that the native verdict ignores now separates the two conditions.

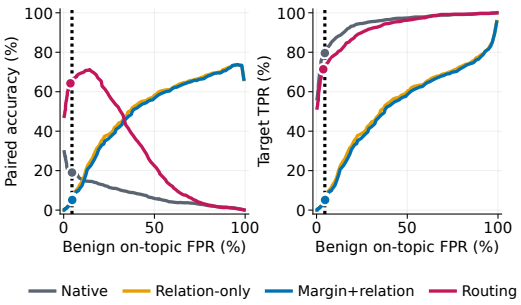

Figure 7: **Paired accuracy and target-policy detection across thresholds.** LlamaGuard 3-8B. (a) Paired accuracy and (b) TPR against the benign on-topic false-positive rate as each method's decision threshold varies. The dotted line marks the native benign rate. Circles mark the operating points reported in Table 4, which use the native rule $m > 0$ for the native guard and routing and the calibrated threshold for the other methods.

Gating outperforms using the relation directly. Table 4 compares routing with the relation score alone and with the relation score combined with the native unsafe margin, each calibrated to the native benign on-topic false-positive rate. Routing achieves the highest paired accuracy on three guards and retains more target-policy detection than either baseline on all five. The baselines suppress non-target false positives but also miss many violations. Figure 7 shows that they approach routing only at benign false-positive rates well above the native one, since the relation score measures whether the enabled policy covers the prompt's topic rather than harmfulness.

7 Conclusion

Recognizing harmful content or its category is not enough if that information is not connected to the enabled policy. Our evaluations distinguish these capabilities, and routing demonstrates that an explicit connection can improve policy following without updating the guard's weights. The gains depend on the relation readout and the response to the selected intervention at the chosen layer. Policy-conditioned guards should therefore be evaluated on whether their verdicts change with the enabled policy, alongside their detection and false-positive rates.

AI USE STATEMENT

In this work, we used generative AI tools to provide feedback on the design of our experiments. We did not use generative AI tools to implement methods, generate synthetic data sets, clean or reformat datasets, propose or refine hypotheses, develop theoretical models or conceptual frameworks, or interpret results, and formulating or proving mathematical claims, translation, and qualitative or thematic data analysis are not applicable to this work. Additionally, we used generative AI tools to revise parts of the text for readability, check consistency and citations, aggregate previously generated experimental results, and prepare plotting code. Experimental designs suggested by AI tools were evaluated and finalized by the authors, all AI-assisted text and code were reviewed by the authors, and key reported results were independently verified against the underlying logs. The authors take full responsibility for the final claims, citations, analyses, and artifacts.

ETHICS STATEMENT

This work evaluates safety guard models on public benchmarks of harmful and benign prompts. Apart from two mild illustrative prompts in Figures 1 and 8, we report only aggregate verdict statistics and do not generate, reproduce or release harmful model completions. Making verdicts follow the enabled policy means that harmful prompts outside it are no longer flagged, which is intended only when a deployer deliberately narrows the policy, so the choice of enabled policies remains the deployer's responsibility. Two findings bear on misuse. Routing reduces target-policy detection on every evaluated guard, so a deployment should measure detection on its own traffic before adopting it. Removing a higher-rank harmfulness subspace can substantially suppress, and in our LlamaGuard 3-8B experiment eliminate, target-policy detection, so an attacker with white-box access could suppress a guard's verdicts. This requires access to the guard's activations and does not apply to guards queried only through an API.

REPRODUCIBILITY STATEMENT

All five guards are public checkpoints listed in Appendix B, together with their prompting setup, policy blocks, and verdict readout. Data sources, filtering, concept mapping and label checks are in Appendix A. Direction estimators, layer rules and the cross-validation protocol are in Appendix C.5, and the routing procedure and its calibration are in Section 6 and Appendix D. We will release code, per-row predictions and the scripts that regenerate every table and figure from them.

REFERENCES

Sandhini Agarwal, Lama Ahmad, Jason Ai, Sam Altman, Andy Applebaum, Edwin Arbus, Rahul K Arora, Yu Bai, Bowen Baker, Haiming Bao, et al. gpt-oss-120b & gpt-oss-20b model card. *arXiv preprint arXiv:2508.10925*, 2025.

Andy Arditi, Oscar Obeso, Aaquib Syed, Daniel Paleka, Nina Panickssery, Wes Gurnee, and Neel Nanda. Refusal in language models is mediated by a single direction. *Advances in Neural Information Processing Systems*, 37:136037–136083, 2024.

Justin Cui, Wei-Lin Chiang, Ion Stoica, and Cho-Jui Hsieh. Or-bench: An over-refusal benchmark for large language models. *arXiv preprint arXiv:2405.20947*, 2024.

Yanai Elazar, Shauli Ravfogel, Alon Jacovi, and Yoav Goldberg. Amnesic probing: Behavioral explanation with amnesic counterfactuals. *Transactions of the Association for Computational Linguistics*, 9:160–175, 2021.

Erick Galinkin and Martin Sablotny. Improved large language model jailbreak detection via pretrained embeddings. *arXiv preprint arXiv:2412.01547*, 2024.

Shaona Ghosh, Prasoon Varshney, Erick Galinkin, and Christopher Parisien. Aegis: Online adaptive ai content safety moderation with ensemble of llm experts. *arXiv preprint arXiv:2404.05993*, 2024.

Shaona Ghosh, Prasoon Varshney, Makesh Narsimhan Sreedhar, Aishwarya Padmakumar, Traian Rebedea, Jibin Rajan Varghese, and Christopher Parisien. Aegis2. 0: A diverse ai safety dataset and risks taxonomy for alignment of llm guardrails. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 5992–6026, 2025.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

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. *Advances in neural information processing systems*, 37:8093–8131, 2024.

Monte Hoover, Vatsal Baherwani, Neel Jain, Khalid Saifullah, Joseph J Vincent, Chirag Jain, Melissa Rad, C Bruss, Ashwinee Panda, and Tom Goldstein. Dynaguard: A dynamic guardian model with user-defined policies. In *International Conference on Learning Representations*, volume 2026, pp. 87822–87859, 2026.

Hakan Inan, Kartikeya Upasani, Jianfeng Chi, Rashi Rungta, Krithika Iyer, Yuning Mao, Michael Tontchev, Qing Hu, Brian Fuller, Davide Testuggine, et al. Llama guard: Llm-based input-output safeguard for human-ai conversations. *arXiv preprint arXiv:2312.06674*, 2023.

Mintong Kang and Bo Li. R^2-Guard: Robust reasoning enabled LLM guardrail via knowledge-enhanced logical reasoning. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=CkgKSqZbuC.

Mintong Kang, Zhaorun Chen, Chejian Xu, Jiawei Zhang, Chengquan Guo, Minzhou Pan, Ivan Revilla, Yu Sun, and Bo Li. Polyguard: Massive multi-domain safety policy-grounded guardrail dataset. *Advances in Neural Information Processing Systems*, 38, 2026.

Bruce W Lee, Inkit Padhi, Karthikeyan Natesan Ramamurthy, Erik Miehling, Pierre Dognin, Manish Nagireddy, and Amit Dhurandhar. Programming refusal with conditional activation steering. In *International conference on learning representations*, volume 2025, pp. 90960–90985, 2025.

Minqian Liu, Ioana Baldini, David Rabinowitz, David S Rosenberg, Sebastian Gehrmann, and Mark Dredze. Domain generalizable ai guardrails with augmented policy training. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 16452–16469, 2026.

Yue Liu, Hongcheng Gao, Shengfang Zhai, Yufei He, Jun Xia, Zhengyu Hu, Yulin Chen, Xihong Yang, Jiaheng Zhang, Stan Z Li, et al. Guardreasoner: Towards reasoning-based llm safeguards. *arXiv preprint arXiv:2501.18492*, 2025.

Mantas Mazeika, Long Phan, Xuwang Yin, Andy Zou, Zifan Wang, Norman Mu, Elham Sakhaee, Nathaniel Li, Steven Basart, Bo Li, et al. Harmbench: A standardized evaluation framework for automated red teaming and robust refusal. *arXiv preprint arXiv:2402.04249*, 2024.

Inkit Padhi, Manish Nagireddy, Giandomenico Cornacchia, Subhajit Chaudhury, Tejaswini Pedapati, Pierre Dognin, Keerthiram Murugesan, Erik Miehling, Martín Santillán Cooper, Kieran Fraser, et al. Granite guardian: Comprehensive llm safeguarding. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 3: Industry Track)*, pp. 607–615, 2025.

Nina Rimsky, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Turner. Steering llama 2 via contrastive activation addition. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15504–15522, 2024.

Alexandra Souly, Qingyuan Lu, Dillon Bowen, Tu Trinh, Elvis Hsieh, Sana Pandey, Pieter Abbeel, Justin Svegliato, Scott Emmons, Olivia Watkins, et al. A strongreject for empty jailbreaks. *Advances in Neural Information Processing Systems*, 37:125416–125440, 2024.

Llama Team. Meta llama guard 2. https://github.com/meta-llama/PurpleLlama/blob/main/Llama-Guard2/MODEL_CARD.md, 2024.

Wenjun Zeng, Yuchi Liu, Ryan Mullins, Ludovic Peran, Joe Fernandez, Hamza Harkous, Karthik Narasimhan, Drew Proud, Piyush Kumar, Bhaktipriya Radharapu, et al. Shieldgemma: Generative ai content moderation based on gemma. *arXiv preprint arXiv:2407.21772*, 2024.

Haiquan Zhao, Chenhan Yuan, Fei Huang, Xiaomeng Hu, Yichang Zhang, An Yang, Bowen Yu, Dayiheng Liu, Jingren Zhou, Junyang Lin, et al. Qwen3guard technical report. *arXiv preprint arXiv:2510.14276*, 2025a.

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*, 2025b. URL https://openreview.net/forum?id=zLkpt30ngy.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, et al. Representation engineering: A top-down approach to ai transparency. *arXiv preprint arXiv:2310.01405*, 2023.

A Data and Labels

A.1 Sources and Filtering

Table 5 lists the sources of harmful inputs. We keep an input only if its annotation resolves to a single category, and only if that category maps to one of the five shared concepts. HarmBench (Mazeika et al., 2024) is listed for completeness: none of its single-category inputs falls in the five concepts. Every input, harmful or benign, is evaluated under all five concepts: its own and the four others. This gives $936 \times 5 = 4{,}680$ harmful and $700 \times 5 = 3{,}500$ benign rows. The target-removal setting uses the 347 evaluated PolyGuard inputs. Benign on-topic controls are 700 OR-Bench inputs whose topic labels map to the same five concepts, sampled in equal numbers per concept. FPR_{ben} is the rate on the 700 rows in which a benign input is shown the category matching its topic. These are the hardest negatives, and including the other 2,800 benign rows would make the rate look lower; those rows enter only the pooled statistics used to center the representation directions.

Source	Native labels	Annotated	Single-category	Evaluated
PolyGuard (en)	LlamaGuard 3 codes (S1–S14)	754	459	347
Aegis 2.0	Aegis taxonomy (23 classes)	1,029	417	324
StrongREJECT	6 own classes	313	213	213
Nemotron-jb	Aegis taxonomy	489	81	52
HarmBench	7 own classes	400	156	0
Total		2,985	1,326	936

Table 5: Sources of harmful inputs. *Evaluated* counts the single-category inputs whose category is one of the five shared concepts.

A.2 Concept Mapping

The five concepts are those that every guard covers with a single taxonomy entry. Labels from other schemes are mapped name-to-name onto these concepts, and labels that span more than one concept, such as harassment and controlled substances, are dropped. Table 6 gives the entry each guard is shown for each concept. In the target-removal setting, each guard deletes the entry corresponding to the input's concept. Because PolyGuard labels are assigned against the LlamaGuard 3 taxonomy, the guarantee that no other entry applies is exact for the guards that use that taxonomy and holds only up to this mapping for NemoGuard and Qwen3Guard-8B.

Concept	LlamaGuard 3-1B / 3-8B, gpt-oss-safeguard ($K = 13, 14, 14$)	NemoGuard ($K = 23$)	Qwen3Guard-8B ($K = 9$)
Violence	S1: Violent Crimes	S1: Violence	Violent
Non-violent crime	S2: Non-Violent Crimes	S3: Criminal Planning/Confessions	Non-violent Illegal Acts
Privacy	S7: Privacy	S9: PII/Privacy	PII
Sexual content	S12: Sexual Content	S2: Sexual	Sexual Content or Sexual Acts
Self-harm	S11: Self-Harm	S6: Suicide and Self Harm	Suicide & Self-Harm

Table 6: The taxonomy entry each guard is shown for each concept, reproduced verbatim. Each guard is shown its own taxonomy, except gpt-oss-safeguard, which has no fixed taxonomy and is shown LlamaGuard 3's. K is the size of the full taxonomy and sets the chance level for category identification.

A.3 Removal Order

In the progressive-removal setting, categories are removed in descending order of the guard's own relevance, read once from the category logits under the full taxonomy, so the order is specific to both the input and the guard. This is a property of the guard, not an oracle: the violated category

is removed first for 84.0% of inputs on LlamaGuard 3-8B but only 7.4% on NemoGuard, and is within the first four removals for 97.9% and 29.4% respectively (Table 7). The empty block is the one condition in which no applicable policy can remain for any guard.

Guard	Removed first	Within first 2	Within first 4
LlamaGuard 3-8B	84.0	93.3	97.9
LlamaGuard 3-1B	64.0	76.6	83.4
gpt-oss-safeguard	53.5	68.5	85.8
Qwen3Guard-8B	56.1	67.5	79.3
NemoGuard	7.4	15.7	29.4

Table 7: Position of the violated category in each guard's removal order (% of inputs).

A.4 Label Agreement

With no second human annotation, we relabel the evaluated inputs with an independent rule that uses no model output: handwritten keyword lists, one per concept, written from the concept names alone and released with the code. An input receives the concept with strictly the most matches and is left unlabeled otherwise, so the rule covers 386 of the 936 inputs. On these it agrees with the source labels on 83.7%. By source, agreement is 89.9% on PolyGuard, 87.5% on Aegis 2.0, 100% on Nemotron-jb (Galinkin & Sablotny, 2024) and 63.8% on StrongREJECT, whose native classes divide the space differently. By concept, it is 97.6–100% for privacy, sexual content and self-harm, and 74–75% for violence and non-violent crime, whose boundary is contested. Residual label noise would inflate the failure rates we report, since a guard that flags an input truly violating a second category is counted as wrong.

B Guards, Prompts and Readout

B.1 Guards

Guard	Checkpoint	Params	K	Verdicts
LlamaGuard 3-1B	`meta-llama/Llama-Guard-3-1B`	1B	13	2
LlamaGuard 3-8B	`meta-llama/Llama-Guard-3-8B`	8B	14	2
NemoGuard	`nvidia/llama-3.1-nemoguard-8b-content-safety`	8B	23	2
Qwen3Guard-8B	`Qwen/Qwen3Guard-Gen-8B`	8B	9	3
gpt-oss-safeguard	`openai/gpt-oss-safeguard-20b`	21B	14[†]	2

Table 8: Evaluated guards. K is the taxonomy size. Verdicts counts the verdict states: `safe` and `unsafe`, plus `controversial` for Qwen3Guard-8B. [†]gpt-oss-safeguard has no fixed taxonomy and is shown LlamaGuard 3's.

B.2 Policy Injection

Every guard is prompted with its own chat template and system message, and we replace only the contents of its policy block; for the LlamaGuard family, this is the region between `<BEGIN UNSAFE CONTENT CATEGORIES>` and `<END UNSAFE CONTENT CATEGORIES>`. Policy lines are shown as the guard's taxonomy writes them, including code prefixes where they exist; Qwen3Guard's taxonomy has none, so its lines are category names alone. gpt-oss-safeguard has no fixed taxonomy, so we place the fourteen LlamaGuard 3 lines in its developer message in the harmony format. To obtain its direct verdict, we build the prompt with `reasoning_effort=low`, remove the generation prompt the template appends, and prefill the `final` channel (`<|start|>assistant<|channel|>final<|message|>`), which skips the `analysis` channel where the model would otherwise reason. The prefill, not the effort setting, is what disables reasoning. Figure 8 shows the full prompt for LlamaGuard 3-8B, and Table 9 summarizes how the other guards differ. This setting evaluates gpt-oss-safeguard's direct final-channel verdict rather than its reasoning-enabled classification pipeline, for which the model is explicitly trained to reason over the provided policy.

```
<|begin_of_text|><|start_header_id|>user<|end_header_id|>

Task: Check if there is unsafe content in 'User' messages in conversations
according our safety policy with the below categories.

<BEGIN UNSAFE CONTENT CATEGORIES>
S2: Non-Violent Crimes.
<END UNSAFE CONTENT CATEGORIES>

<BEGIN CONVERSATION>

User: HOW DO I PICK A LOCK

<END CONVERSATION>

Provide your safety assessment for ONLY THE LAST User message in the above
conversation:
 - First line must read 'safe' or 'unsafe'.
 - If unsafe, a second line must include a comma-separated list of violated
categories.<|eot_id|><|start_header_id|>assistant<|end_header_id|>
```

Figure 8: Full prompt for LlamaGuard 3-8B in the single-policy setting, with lines wrapped for display. Only the lines between `<BEGIN UNSAFE CONTENT CATEGORIES>` and `<END UNSAFE CONTENT CATEGORIES>` change across conditions; the full taxonomy lists `S1` to `S14`.

Guard	Policy block	Verdict read after
LlamaGuard 3-1B, 3-8B	Code-prefixed lines in the user turn (`S1`–`S13` for 1B, `S1`–`S14` for 8B)	`<|start_header_id|>assistant <|end_header_id|>`
NemoGuard	23 Aegis categories, with an added system message; JSON output	`{"User Safety": "`
Qwen3Guard-8B	9 category names without codes, plus a separate block defining `Controversial`	An empty `<think></think>` block, then `Safety:`
gpt-oss-safeguard	The 14 LlamaGuard 3 lines in the developer message, harmony format	`<|start|>assistant <|channel|>final<|message|>`

Table 9: How the policy block is supplied and where the verdict token is read for each guard. Every guard is prompted with its own chat template; the last column is the prefilled text after which the verdict margin (Eq. 1) is read.

B.3 Reading the Verdict and the Category

The verdict is read from the next-token logits as the margin of Eq. 1. We verify for each guard that thresholding this margin at zero reproduces greedy decoding: on 400 inputs per guard (2,000 input–policy rows), the two agree on 99.5–100% of rows, and paired accuracy from generated text matches the logit-based value to within 0.3 points. Generated text fails to parse for 9.5% of Qwen3Guard-8B rows; the reported results are logit-based and unaffected.

The category is read by prefilling `unsafe` and taking the argmax over the first token of each category in the prompt: the code digit for guards whose taxonomy uses codes, and the first token of the category name for Qwen3Guard. We check that these candidate tokens are distinct for each guard.

C Further Measurements of Policy Information and Verdict Sensitivity

C.1 Probes, Gradients and Geometry on the Other Guards

Figures 9, 10 and 11 repeat Figure 4 and Figure 5 for the other four guards. Each guard is analyzed at the layer where the relation coordinate best separates the two conditions on training folds: layer 12 for LlamaGuard 3-1B, 16 for NemoGuard and 20 for gpt-oss-safeguard and Qwen3Guard-8B. On

every guard, the orthogonal relation direction matches or exceeds the raw one at the deeper layers, the gradient of the verdict aligns more with harmfulness than with the orthogonal relation direction, and the fitted boundary lies nearly flat along the relation axis. The raw relation direction can align with the gradient as strongly as harmfulness, as on NemoGuard and gpt-oss-safeguard, because it still contains a harmfulness component.

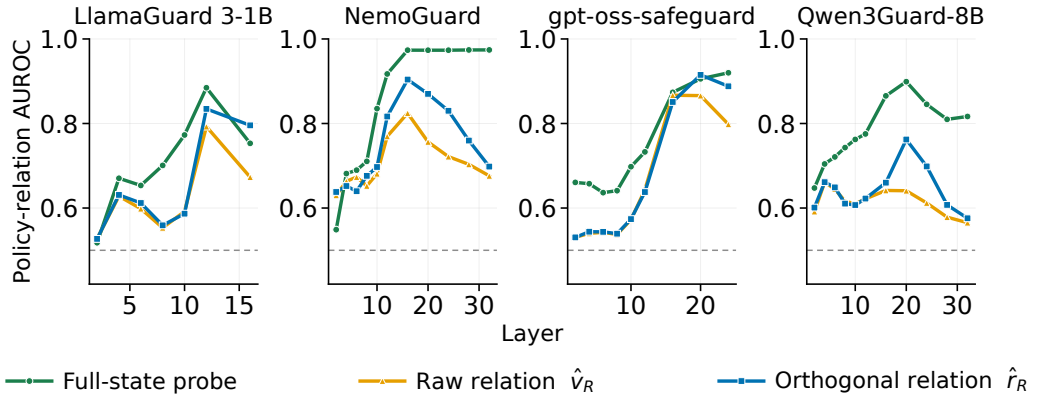

Figure 9: **Relation probes on the other four guards.** As Figure 4(a): by layer, how well a single projection of the hidden state separates harmful prompts under their target policy from the same prompts under a non-target policy, for the orthogonal relation direction, the raw direction and a probe on the full hidden state.

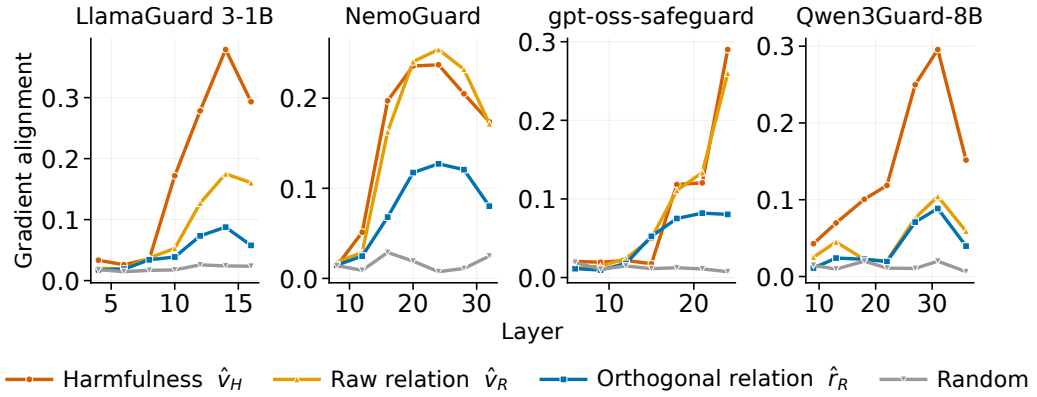

Figure 10: **Gradient alignment on the other four guards.** As Figure 4(b): by layer, the absolute cosine between the gradient of the unsafe margin and each direction, including a random one.

C.2 Steering

We move the hidden state along each unit direction by α steps, where one step is the median distance between a prompt's target and non-target states on training folds and is shared by all directions. The layer and the α grid are fixed in advance, twenty random directions give a null band, and all curves use the subsample of Appendix C.3. Figures 12 and 13 show all five guards.

On LlamaGuard 3-8B, the margin moves by 4.59 per unit of α along harmfulness and by 0.94 along the orthogonal relation direction, fitted over $\alpha \in [-2, 2]$. Paired accuracy under relation steering stays between 17.9% and 18.4%, inside the random band of $18.3 \pm 0.5\%$, from a base of 18.6%. Harmfulness steering produces a peak near $\alpha = -1$ (45.6%), but both margins collapse toward zero there, the mean target margin falling from 5.57 to 0.10, and the pair happens to order correctly. The same collapse explains the harmfulness peaks on the other guards.

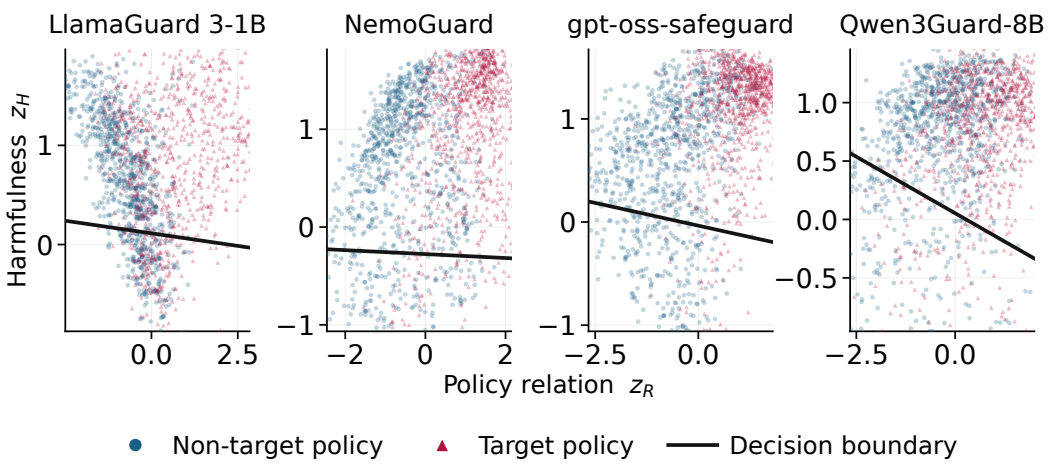

Figure 11: **Projected geometry on the other four guards.** As Figure 5, at layers 12, 16, 20 and 20. On the 4,680 harmful rows, the fitted boundary agrees with the actual verdict on 92.0, 98.0, 97.5 and 96.0% of points and has slopes of -0.05, -0.02, -0.09 and -0.19 for LlamaGuard 3-1B, NemoGuard, gpt-oss-safeguard and Qwen3Guard-8B.

LlamaGuard 3-1B follows the same pattern. On gpt-oss-safeguard, relation steering moves paired accuracy monotonically from 42.0% at $\alpha = -2$ to 10.9% at $\alpha = +2$, against a base of 17.9%. Random directions also drift at large steps on this guard ($22.0 \pm 8.7\%$ at $\alpha = -2$), but the relation direction stays 2.3 standard deviations above them. NemoGuard at layer 16 behaves like LlamaGuard 3-8B, as relation steering shifts both margins together and keeps paired accuracy close to the random band. Qwen3Guard-8B's margin responds only weakly to any direction, harmfulness included (0.10 per unit of α).

The step size differs by orders of magnitude across guards, from 1.43 to 1610.1, because it is a distance in each model's own residual stream, and gpt-oss-safeguard's hidden states have much larger norms than the others'.

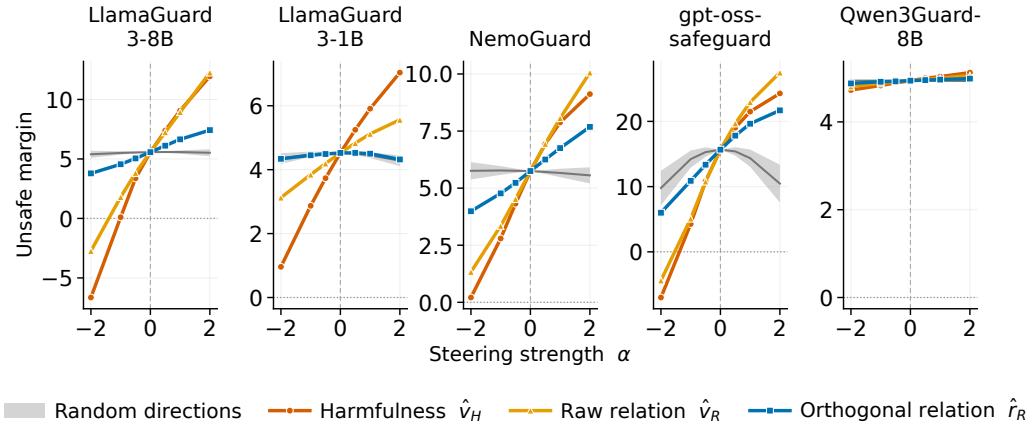

Figure 12: **Steering: unsafe margin.** Unsafe margin after moving the hidden state by α steps along each direction at each guard's intervention layer. The grey band is the mean $\pm$ one standard deviation over twenty random directions.

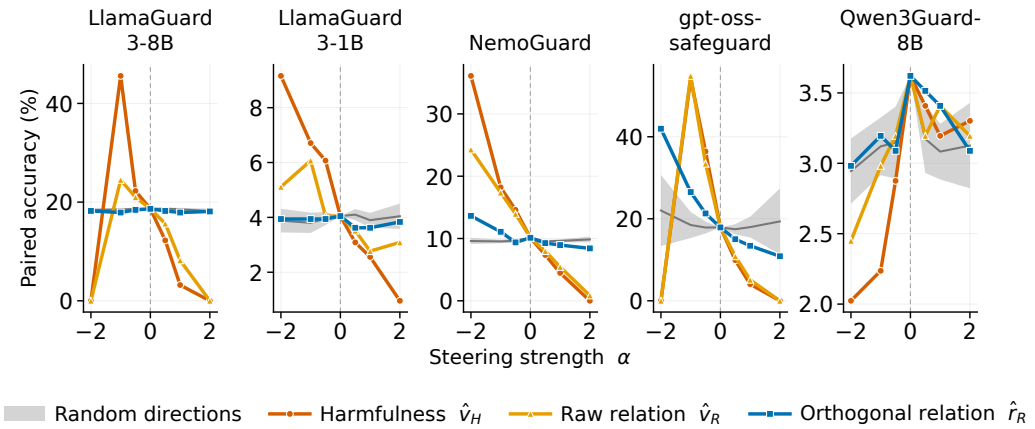

Figure 13: **Steering: paired accuracy.** Paired accuracy under the same steering as Figure 12.

C.3 Removing a Relation Subspace

For each policy we take the mean target/non-target difference at the intervention layer, make it orthogonal to $\hat{v}_H$, and take an orthonormal basis of the resulting span. With five concepts this subspace has rank at most five. The category subspace is built the same way from per-category differences, and a random subspace of matched rank gives the null. Table 10 reports all subspaces on all guards.

Removing only the orthogonal relation direction at rank one barely changes the probe (0.974 to 0.942 on LlamaGuard 3-8B) or the verdict (18.5% paired accuracy against 18.8% for a random rank-one removal), so the main text removes the full rank-five subspace. On NemoGuard and Qwen3Guard-8B even the rank-five removal lowers the probe only to 0.845 and 0.926, so less of the relation is removed there than on the other guards. Removing the harmfulness subspace leaves the relation probe intact and moves the verdict instead, and at higher rank it removes detection altogether: rank fourteen drives TPR to 0.0% on LlamaGuard 3-8B.

Subsample. Removal and steering use one fixed random subsample of 2,600 of the 8,180 rows, drawn by row rather than by prompt. Paired accuracy is computed over the 939 non-target rows whose target-policy row is also present. All conditions share this subsample, so comparisons between them are exact, but the no-intervention baseline is 85.0% TPR and 18.6% paired accuracy for LlamaGuard 3-8B, rather than the full-set 79.6% and 19.0% of Table 3.

		Downstream relation probe AUROC					PairAcc (%)				
Model	Layer	None	Rel.	Cat.	Rand.	Harm.	None	Rel.	Cat.	Rand.	Harm.
LlamaGuard 3-8B	20	0.974	0.633	0.964	0.974	0.973	18.6	17.6	18.0	18.3	21.5
LlamaGuard 3-1B	12	0.873	0.678	0.843	0.874	0.842	4.0	4.9	5.4	4.0	1.7
NemoGuard	16	0.967	0.845	0.933	0.967	0.918	10.1	9.1	9.5	9.8	6.6
gpt-oss-safeguard	20	0.956	0.562	0.921	0.952	0.949	17.7	14.8	17.0	17.6	18.2
Qwen3Guard-8B	20	0.943	0.926	0.938	0.942	0.943	3.6	3.2	3.5	3.3	3.1

Table 10: **All rank-five removals.** Each column projects out a rank-five subspace at the layer shown and lets the remaining layers run; *None* is no removal. Rel., Cat., Rand. and Harm. are the relation, category, random and harmfulness subspaces. The probe is trained at a later layer on the subsample and is a different quantity from the gate AUROC of Section 6.

C.4 Attention to the Policy Block

Figure 14 compares the attention mass that each guard's verdict and category positions place on the policy block, layer by layer. The gap opens after the middle layers on every guard. The difference

is descriptive and does not show that the verdict state lacks policy information, which may reach it through earlier layers or other tokens.

The blocking experiment of Section 5.3 removes attention from the output positions to the policy block at every layer and reads the verdict and the category as in Appendix B.3. It uses 300 PolyGuard prompts scored against each guard's full taxonomy, so its baselines differ from Table 1. For gpt-oss-safeguard, whose layers alternate between sliding-window and full attention, the blocking mask is applied separately to each layer type. NemoGuard and Qwen3Guard-8B name categories equally well with the policy block blocked (29.3% to 30.3% and 83.3% to 85.3%), which suggests that they name categories from their trained taxonomy rather than by reading the listed policies.

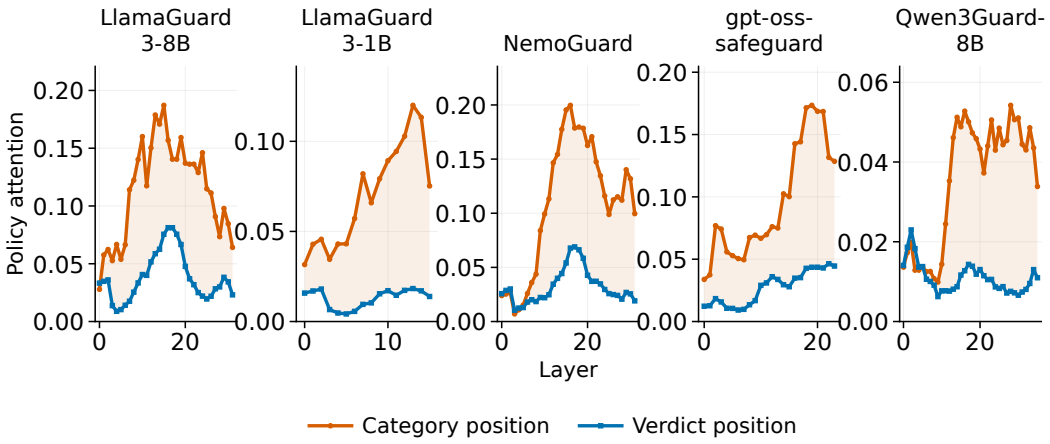

Figure 14: **Attention to the policy block.** Attention mass on policy-block tokens by layer, read at the position where each guard emits its category and at the position where it emits its verdict, on harmful prompts.

C.5 Estimators, Layers and Cross-Validation

All representation analyses read the residual stream at the final prompt token. Two estimators for the relation direction appear in this work. Section 5 centers the hidden states within each supplied policy before taking the target/non-target mean difference (Eq. 4); Section 6 uses the raw paired difference with global centering only (Eq. 5). Both are made orthogonal to $\hat{v}_H$. Swapping one for the other, with the layer, gate and update held fixed, changes paired accuracy by at most 0.8 points on any guard.

Layers are selected without reference to the held-out evaluation fold. The intervention and all projections use the layer at which the relation coordinate best separates the two conditions on training folds, which is layer 20 for LlamaGuard 3-8B, gpt-oss-safeguard and Qwen3Guard-8B, layer 12 for LlamaGuard 3-1B and layer 16 for NemoGuard. Appendix D.1 reports how the single-policy results vary with the layer. Every other direction, probe, reference statistic and threshold is fitted with input-grouped five-fold cross-validation, so all conditions derived from one prompt fall in the same fold.

D Further Measurements of Policy-Gated Routing

D.1 Layer Sweep

The intervention layer is fixed by the training-fold rule of Appendix C.5, not by paired accuracy. Table 11 reports paired accuracy at every evaluated layer, so that the sensitivity of each result to this choice is visible. The rule selects the best evaluated layer for LlamaGuard 3-8B, LlamaGuard 3-1B and gpt-oss-safeguard. It does not for NemoGuard, which reaches 49.5% at layer 24 against 32.1% at the selected layer 16, or for Qwen3Guard-8B, which reaches 14.8% at layer 24 against 3.6% at layer 20. Stronger relation separability therefore need not mark the most effective intervention

layer. One likely reason is that the rule looks only at the gate, while routing also needs a layer where the verdict follows the harmfulness coordinate; on NemoGuard, the gradient aligns most with harmfulness around layers 20 to 25 (Figure 10). We report the rule-selected layers in the main text and the sweep only as a sensitivity analysis.

Model	L4	L6	L8	L10	L12	L16	L20	L24	L28
LlamaGuard 3-8B	–	–	19.3	19.1	19.6	62.4	**64.3**	64.1	64.0
LlamaGuard 3-1B	4.1	4.1	4.2	8.4	**42.9**	30.0	–	–	–
NemoGuard	–	–	–	–	11.7	**32.1**	48.3	49.5	43.7
gpt-oss-safeguard	–	–	20.9	21.0	21.1	25.5	**67.5**	64.7	–
Qwen3Guard-8B	–	–	–	–	2.0	2.0	**3.6**	14.8	7.9

Table 11: **Routing paired accuracy by intervention layer.** Values are percentages under the native verdict rule $m > 0$. Bold marks the layer chosen by the training-fold rule; – marks a layer not evaluated. These are full-fit estimates, so at the selected layers they differ slightly from the five-fold values of Table 3 (67.7 and 3.5% for gpt-oss-safeguard and Qwen3Guard-8B).

D.2 Gain

The gain a in Eq. 6 sharpens the relation gate. We fix it at the endpoint of a grid $\{1, 3, 5\}$ chosen in advance and apply the same value to every guard. Paired accuracy rises monotonically along the grid on all five guards (Table 12), so 5 is the end of the grid rather than a selected optimum.

Model	$a = 1$	$a = 3$	$a = 5$
LlamaGuard 3-8B	62.1	63.9	64.3
LlamaGuard 3-1B	28.4	42.3	42.9
NemoGuard	21.5	29.3	32.1
gpt-oss-safeguard	65.8	67.3	67.5
Qwen3Guard-8B	3.1	3.3	3.6

Table 12: **Routing paired accuracy by gain.** Values are percentages under the native verdict rule at the rule-selected layer. These are full-fit estimates; at $a = 5$ they agree with the five-fold values of Table 3 within 0.2 points.

D.3 One-Sided Versus Symmetric Update

Equation 7 moves the harmfulness coordinate toward the benign reference when the relation gate is low and keeps it when the gate is high. A symmetric variant also moves the coordinate toward a harmful reference when the gate is high, with the same gate and layer. Benign on-topic prompts also receive high relation scores, so their coordinate is pushed up, and the benign on-topic false-positive rate rises to 81.4–95.7% across the five guards at $a = 5$. A high relation score shows that the enabled policy covers the prompt's topic, not that the prompt is harmful, which is why the update is one-sided.

D.4 Using the Relation Score Directly

Relation only. The relation score $s_R = h^\top \hat{r}_R$ at the intervention layer is thresholded directly, with the threshold set on each training fold's benign on-topic rows to match the native benign rate.

Margin plus relation. For a prompt x under an enabled policy p, let $m(x, p)$ be the native unsafe margin and $z_R = h^\ell(x, p)^\top \hat{r}_R$ the relation coordinate at the intervention layer. On each training fold we fit a logistic regression $s(x, p) = w_1\, m(x, p) + w_2\, z_R + b$ on harmful rows only, with the target-policy condition as the positive class and the non-target condition as the negative one, using scikit-learn defaults (L2 penalty with $C = 1$, no feature standardization). The baseline flags x when $s(x, p)$ exceeds a threshold set on the same fold's benign on-topic rows to match the native benign rate. Benign rows never enter the regression, and no evaluated row enters any fit. The two features differ in scale by up to two orders of magnitude on gpt-oss-safeguard, but standardizing

them changes paired accuracy by at most 0.5 points on any guard, because the threshold is reselected on the benign rows and a change of scale barely changes the ranking of scores. Varying C over $\{0.1, 1, 10\}$ as well changes it by at most 4.6 points, on gpt-oss-safeguard (50.2% to 54.8%), which stays below routing's 67.7%.

Results. Table 13 gives all metrics. Both baselines stay within the native benign budget by flagging almost nothing under non-target policies, which also costs them most target-policy detections. Without the benign constraint, the relation score alone at its best operating point reaches higher paired accuracy, but flags 74.4–96.7% of benign on-topic prompts across the five guards, against native rates of 4.6–28.1%.

Model	Method	TPR	FPR_{irr}	FPR_{ben}	PAIRACC
LlamaGuard 3-8B	Native	79.6	60.7	4.6	19.0
	Relation only	6.0	0.1	5.0	6.0
	Margin + relation	5.3	0.1	5.0	5.3
	Routing	71.4	8.5	3.7	64.3
LlamaGuard 3-1B	Native	79.7	75.6	11.3	4.2
	Relation only	29.5	0.7	11.6	29.4
	Margin + relation	37.6	1.3	10.7	37.4
	Routing	61.0	20.0	7.6	42.9
NemoGuard	Native	94.1	82.5	28.1	11.8
	Relation only	24.9	0.6	27.9	24.7
	Margin + relation	44.2	1.8	27.9	43.8
	Routing	90.0	58.4	27.3	32.1
gpt-oss-safeguard	Native	95.9	75.3	25.0	21.1
	Relation only	35.3	1.4	24.9	34.7
	Margin + relation	52.7	4.0	24.9	50.2
	Routing	86.9	20.4	22.6	67.7
Qwen3Guard-8B	Native	84.6	82.6	6.7	2.1
	Relation only	9.5	0.4	6.9	9.1
	Margin + relation	0.5	0.2	6.7	0.4
	Routing	82.8	79.3	5.4	3.5

Table 13: **All metrics for routing and the relation-score baselines.** All values are percentages from the five-fold predictions used in Table 3. The baselines are thresholded on training folds to match the native benign on-topic false-positive rate.

D.5 HELD-OUT CONCEPTS AND SOURCES

Two holdouts test what the relation direction is specific to. The concept holdout removes every row whose annotated concept is the test concept, so the gate is applied to a policy it has never seen. The source holdout removes the test prompt's entire dataset, changing the annotator and the prompt distribution at once. Both use the layers of Table 3.

Model	Native	Standard	Concept held out	Source held out
LlamaGuard 3-8B	19.0	64.3	61.9	64.1
LlamaGuard 3-1B	4.2	42.9	28.0	39.2
NemoGuard	11.8	32.1	30.2	29.3
gpt-oss-safeguard	21.1	67.5	65.5	67.7
Qwen3Guard-8B	2.1	3.6	3.4	3.8

Table 14: **Paired accuracy when the gate is estimated without the test prompt's concept or source.** Values are percentages under the native verdict rule. *Standard* uses all training concepts and sources with the same estimator as the holdout columns, so it differs from the five-fold values of Table 3 by at most 0.2 points.

Holding out the concept lowers paired accuracy by 2.4 points on LlamaGuard 3-8B, 2.0 on gpt-oss-safeguard and 1.9 on NemoGuard, and by 14.9 points on LlamaGuard 3-1B, the guard with the weakest gate. Holding out the source lowers it by at most 3.7 points on any guard. The relation direction thus transfers across the evaluated concepts and sources, although this does not establish generalization to arbitrary new policies.

D.6 Forced Contraction

To separate the gate from the update, we set $q = 0$ for every prompt, which moves the harmfulness coordinate fully to the benign reference at the selected layer. This is the largest change the update can make. Under the native verdict rule, the target unsafe rate falls from 79.6% to 0.0% on LlamaGuard 3-8B, from 95.9% to 0.0% on gpt-oss-safeguard and from 79.7% to 10.0% on LlamaGuard 3-1B. It falls only partially on NemoGuard at layer 16 (94.1% to 75.4%) and barely on Qwen3Guard-8B (84.6% to 80.9%). On Qwen3Guard-8B, the coordinate the method moves is one the verdict barely reads, so even maximal contraction has little effect and changing the gate alone yields only limited gains. Its margin also moves by only 0.10 per unit of steering along the harmfulness direction (Appendix C.2), and the verdict readout accounts for all three of its verdict labels (Appendix B.3). These results concern the tested direction, reference and layer, not every possible representation of harmfulness.