BEYOND DECISION FLIPS: AN EMPIRICAL STUDY OF LLM SAFETY JUDGMENTS

Anonymous authors*
Paper under double-blind review

ABSTRACT

Adding noise to a language model's hidden states shows at which layers its safety judgments change, but not why. We study these responses by dividing the same noise energy among tokens in two ways, reversing the noise sign, adding noise only at some token positions, and predicting responses on new inputs. How the noise is divided among tokens determines whether the response depends on the gradient of the clean decision score, and a noise vector and its negative can move a judgment in the same direction. Position interventions link this response to an uneven output change, in which the differences between inputs' outputs shrink much more than the average output. On Qwen2.5-7B, noise before the final prompt position scales these differences to about 0.55 of their clean size, compared with 0.87–0.93 when only the final position is perturbed. At one responsive layer, a model of this change predicts held-out decision margins with RMSE 3.31, versus 7.16 for uniform scaling and 3.28 for a fit made directly to the margins. Models and layers differ in whether a large response comes with this shrinkage, and labeled outcomes show whether the changes help or hurt accuracy. The tests link a layer's response to the output changes behind it.

1 INTRODUCTION

Adding noise to a language model's hidden states at one layer and counting how many safety judgments flip measures that layer's response; doing so for every layer gives a response profile. A layer with many flips, a response peak, can arise in several ways. The noise may move hidden states along the direction that most changes the decision score, it may change earlier token states that later layers pass on to the answer position, or it may change the output scores in the same way for all inputs. All three can produce similar flip rates, yet they have different explanations and different consequences for safety.

Refusal directions and contrastive activation addition find directions in hidden-state space that change outputs (Arditi et al., 2024; Panickssery et al., 2024). Latent attacks test safety under hidden-state changes (Gu et al., 2025), and layer selection methods find promising intervention sites (Li et al., 2025; Sun et al., 2025). Steering along random directions can also cause some of the same behavioral shifts as learned directions (Malla et al., 2026). We ask how the noise allocation, the way a fixed noise energy is divided among tokens, changes how much the response depends on the gradient of the decision score. We also ask which token positions carry the response and whether changes in the non-class outputs, the logits of tokens other than the two class answers, predict each input's safety judgment. The answers link the applied noise to the resulting change in the decision.

Figure 1 summarizes the main results. A noise vector and its negative can move a judgment the same way (panel a). At responsive Qwen7B layers, noise shrinks the differences between inputs' outputs much more than it changes their average (panel b). Noise at earlier prompt positions produces much of this change without directly perturbing the final position, from which the answer is predicted. A model of this uneven change, fitted on some inputs, predicts decision margins on other inputs and noise (panel c), so the decision response is part of a broader output change. Comparing models and layers shows when this output change accounts for the decision response. Opposite noise signs test whether a response follows the noise direction, and noise at only some positions (position masks) tests how the response reaches the answer. Prediction across inputs then tests whether the output changes follow a rule shared by all inputs or only produce a large average response.

*Code: `https://anonymous.4open.science/r/Boundary-Profiling-DD46/`

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.

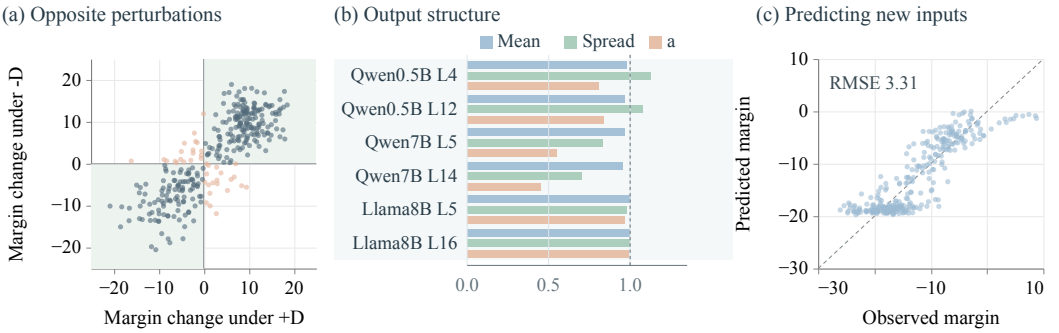

Figure 1: Noise responses have structure beyond decision flips. (a) At Qwen7B layer 14, adding a noise tensor and its negative can move the margin in the same direction (shaded quadrants). (b) Under global noise at all six sites, relative to clean outputs: the RMS of the mean output (Mean), the spread across inputs (Spread), and the clean-pattern coefficient a of Equation 3, which measures how much of the clean differences between inputs remains. (c) Held-out margin prediction at Qwen7B layer 14. Panels (a,c) each show 320 input/direction pairs from 160 inputs; (b) uses the inputs of the position experiments.

2 Study Design

Decision measures and scope. We read the model's decisions in three ways. First, the model generates a harmful/benign judgment for ToxicChat-derived inputs, and we measure the flip rate at each layer. Second, experiments on class scores use each model's official chat template and read the next-token logits of the two class answers, which give a continuous score. Third, evaluations of prompt formatting measure refusal and utility during generation on harmful and benign requests from AdvBench, JailbreakBench, OR-Bench, and XSTest (Zou et al., 2023b; Cui et al., 2024; Röttger et al., 2024). The first two concern toxicity judgments.

Margins and paired responses. For input x, let $H_\ell(x)$ be the clean (noise-free) output of transformer block ℓ at all prompt tokens and D an additive perturbation. With the input and prompt fixed, we compare the decisions before and after replacing H_ℓ by $H_\ell + D$. Flip rate is the fraction of changed binary decisions. For class scores, the decision margin $m_0(x)$ is the clean logit of the toxic answer minus that of the non-toxic answer; its sign is the decision. The margins $m_+(x)$ and $m_-(x)$ add the same sampled perturbation with opposite signs, and we call the two a paired response. We measure the response to noise as margin changes and as sign changes (decision flips). All margins are kept, even when the top token over the whole vocabulary is not a class answer; how often this happens is recorded separately.

Noise allocation and position. We compare two Gaussian noise types with the same total energy (squared Frobenius norm) that divide it differently among tokens; we call this division the *allocation*. *Global* noise uses one scale for all tokens, set by the spread of the whole clean block output. *Activation-weighted* noise gives larger noise to tokens whose clean hidden states have larger RMS and is rescaled to the same total norm. Writing r_t for the clean activation RMS at position t and Z_t for its Gaussian draw, the weighted tensor has entries $R_t = r_t Z_t$ and is scaled as $D_A = \|D_G\|_F R / \|R\|_F$, where D_G is the global tensor. A control shuffles the weights across positions to test whether each weight must stay with its own token. Under global noise, a state with a small clean magnitude moves more relative to its own size, whereas activation weighting accounts for this local scale. We therefore test which noise directions each response depends on, instead of comparing flip counts alone. Separate position experiments split an original noise tensor into its parts at earlier positions and at the final position without rescaling either part, so we can trace how the original noise produces its response.

Models and input groups. Layer profiles of generated judgments are repeated on five non-overlapping blocks of 500 ToxicChat examples with Qwen2.5-0.5B-Instruct. Comparisons of components and model sizes include Qwen models up to 72B and Mistral7B. The direction, position,

and prediction experiments use Qwen2.5-0.5B, Qwen2.5-7B, and Llama3.1-8B-Instruct. Their 80 development inputs serve the analyses across all layers and a 20-input pilot; a further direction experiment uses the other 60 Qwen7B inputs. Output prediction uses 160 inputs with cross-fitting, so each input is predicted from a fit on other inputs. The position experiments use 40 Qwen7B inputs for an initial analysis and another 80 for the later comparison; the other two models use 40 each. Appendix B gives the overlaps between input groups, the fitting splits, and every evaluated condition.

Controlled comparisons and uncertainty. The direction, position, and prediction experiments use noise magnitude 0.025, two sampled noise directions, each with both signs, and preselected shallow and middle layers: 4/12 for Qwen0.5B, 5/14 for Qwen7B, and 5/16 for Llama8B. We call each of these six model–layer pairs a site. Each comparison between two interventions uses the same inputs. Bootstrap intervals resample inputs and keep all directions and signs of a sampled input together. For the model-size comparison with generated judgments, uncertainty over nine combinations of layers and magnitudes is reported separately. Human reviewers check whether changes in generated responses change the judgment or only the wording, with overall agreement of 95.5% on 200 reviewed pairs (Appendix I).

3 Effects of Noise Allocation and Sign

3.1 Allocation determines dependence on the decision gradient

Matching the total energy does not fix how far each token's state moves relative to its own size. We ask whether this only rescales the response or changes which noise directions it depends on. We call the gradient of the clean margin with respect to a token's hidden state the *clean decision direction*, the direction in which a small change most increases the margin. At each token, we remove the noise component along it and rescale the rest to the token's original noise norm, so allocation and energy stay the same. The retained response, the fraction of the mean absolute margin change that remains, shows how much the response depends on this direction.

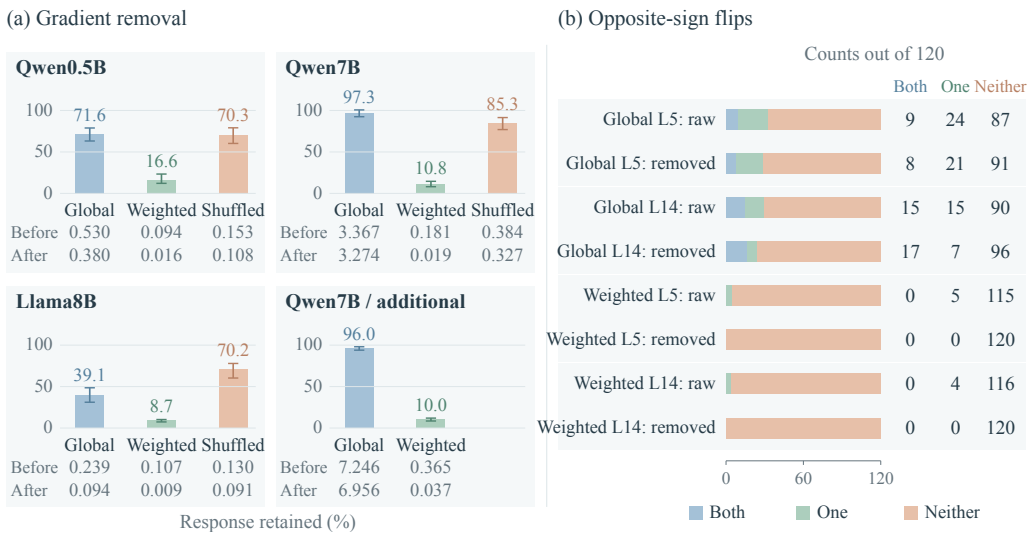

Figure 2: Allocation changes how much the response depends on the clean decision direction. (a) Retained response after removing the noise component along the clean margin gradient, with 95% input-bootstrap intervals; rows below give mean absolute responses before and after removal. The three model panels use 20 inputs each; the additional panel uses 60 other Qwen7B inputs. (b) Number of noise pairs on those additional inputs in which both signs, one sign, or neither flips the clean decision; 120 pairs per row. Under global noise, flips under both signs survive gradient removal.

All three models and the additional inputs show the same contrast (Figure 2a). On the additional 60 inputs, global noise retains 96.0% of its mean absolute margin change, whereas activation-weighted

noise retains 10.0%. The mean absolute changes fall from 7.246 to 6.956 and from 0.365 to 0.037, respectively. In the three-model pilot, shuffled weights retain much more of the response than weights that stay with their own tokens. The result thus depends on which weight goes with which position, and not merely on the noise being uneven across tokens.

The two allocations therefore differ in which noise directions their responses depend on. Under activation weighting, the response comes mostly from the noise component along the clean decision direction. Under global noise, most of the response remains when this component is removed, so it comes from the other directions. The removal takes out both the linear and the nonlinear effects of that component. Two tests with the same energy can therefore measure different things, because the allocation changes which parts of the noise produce the response.

3.2 A NOISE VECTOR AND ITS NEGATIVE CAN FLIP THE SAME DECISION

If the margin responded linearly to the noise, reversing the noise would exactly reverse the margin change. To separate this sign-dependent change from a change shared by both signs, we split each paired response into

$$E(x) = \frac{m_+(x) + m_-(x)}{2} - m_0(x), \qquad O(x) = \frac{m_+(x) - m_-(x)}{2}. \tag{1}$$

Here E is the *shared shift*, the average margin change under the two signs, and O is the *sign-dependent change*, which reverses when the noise is reversed. The two margin changes are $E + O$ and $E - O$. Thus both signs move the margin the same way when $|E| > |O|$. We check whether the shared shift is large enough to change decisions and whether it follows the meaning of the class rather than the letter of the answer.

On the additional inputs, both signs flip the clean decision in 24 of 240 pairs under global noise, and 25 pairs still do so after the clean gradient component is removed (Figure 2b). Reversing the noise thus does not undo the response, and an account in which the margin simply follows the direction of the noise is incomplete. Activation-weighted noise never flips a decision under both signs. Among the 25 global pairs after removal, seven correct the clean decision and 18 introduce errors, so the shared shift can both correct and cause errors.

Swapping the answer letters from A=toxic/B=non-toxic to the reverse keeps the direction of the shared shift for all 20 checked inputs at Qwen7B layer 14, with rank correlation 0.961; at layer 5, 15 of 20 agree. A fixed preference for one answer letter thus cannot explain the layer-14 result. We next test whether a change in the output distribution shared across inputs explains this shift.

4 CHANGES IN THE MEAN OUTPUT AND IN INPUT DIFFERENCES

The shared shift suggests that noise of either sign changes the output scores in a similar way. Suppose the noise keeps the average output over inputs largely intact but shrinks each input's difference from that average. Each input's score then moves toward the average, in a direction set by the side of the average it starts from. If this account is right, the interventions that produce the shared shift should shrink the differences between inputs much more than they change the average output, and changes estimated on some inputs should predict the responses of others.

4.1 SEPARATING THE MEAN FROM EACH INPUT'S DEVIATION

For each input, we subtract the mean over the vocabulary from its next-token logits to obtain $z(x)$. Over a set of inputs, write

$$z(x) = \mu + u(x), \qquad \mu = \frac{1}{N}\sum_x z(x), \qquad u(x) = z(x) - \mu. \tag{2}$$

The *mean output* μ is the average of these logit vectors for the task and input set, and $u(x)$ is input x's *deviation* from it, which records what is specific to that input. We call the clean deviations of all inputs the *clean pattern*. We model the perturbed output as

$$z_D(x) = \mu_D + a\,u(x) + e(x), \tag{3}$$

where μ_D is the perturbed mean output and a is the *clean-pattern coefficient*, the least-squares slope of the perturbed deviations on the clean ones, fitted on vocabulary entries other than the class answers and special tokens; $e(x)$ is the variation this slope leaves unexplained. A value $a = 1$ keeps the differences between inputs at full size, and $a < 1$ shrinks them, which we call *attenuation*. The residual $e(x)$ can add new differences between inputs, so the total spread across inputs can grow even when $a < 1$.

Subtracting the vocabulary mean removes a constant shift of all logits, which cannot change probabilities. Each entry of μ is a token's mean score across inputs, and $u(x)$ records how much input x favors that token relative to the mean.

As an example, suppose two clean margins are -0.5 and 2.5, with mean 1. Keeping the mean fixed and setting $a = 0.5$ gives 0.25 and 1.75. Both scores move toward the mean, and the first decision flips. If both noise signs cause this change, the first input flips under both signs. The same change moves the two margins in opposite directions because they start on different sides of the mean.

4.2 What the account predicts

Taking the difference between the two class logits in Equation 3 gives the margin change

$$m_D(x) - m_0(x) \approx (\bar{m}_D - \bar{m}) + (a - 1)[m_0(x) - \bar{m}] + e_m(x), \tag{4}$$

where $\bar{m}$ and $\bar{m}_D$ are the mean clean and perturbed margins and $e_m(x)$ is the remaining change. The first term moves all margins together. The second depends on each input: when $a < 1$, inputs above the clean mean move downward relative to the new mean, and those below it move upward. If both noise signs produce similar changes in these terms, the changes add to the shared shift E in Equation 1 rather than to O. One change in the output can thus move the margin the same way under both signs, by a different amount for each input.

Three comparisons separate this account from simpler explanations. Uniform scaling (one factor for all logits) would scale the mean output and the deviations equally; separate estimates of μ_D and a test this restriction. A shared offset, which adds the same vector to every input's output, would keep the clean pattern with $a = 1$; held-out prediction tests whether also changing its scale adds information. Finally, a map fitted to the class logits could describe the two answer logits alone; because a is estimated outside those entries, a good margin prediction from a shows that a broader output change also predicts the decision. We run these comparisons at each selected layer.

5 Testing the Explanation through Positions and Prediction

5.1 Noise at earlier positions produces most of the change

Separating earlier and final positions. We split the same noise tensor into two parts, $D = D_{\text{early}} + D_{\text{final}}$, and apply each part separately. D_{final} holds the noise at the final position, the last prompt token after the chat template, from which the answer is predicted; D_{early} holds the noise at all earlier positions. With D_{early} alone, the final position is unperturbed at the intervention layer, but later layers can still pass changes to it. Neither part is rescaled. A part's *retention* is its mean absolute margin change divided by that of the full noise.

Figure 3 shows that a large response does not require perturbing the final position directly. On the additional 80 Qwen7B inputs, noise at earlier positions retains 97.1% and 83.7% of the full response at layers 5 and 14, whereas noise at the final position retains 12.0% and 21.8%. At the other four sites, retention for earlier positions ranges from 80.8% to 99.0%. Earlier positions hold about 98.9% of the global noise energy; the retention values describe each part as it occurs in the original noise, and the signed interaction (the full change minus the sum of the two parts' changes; Appendix B) shows how the two parts combine.

In Qwen7B, earlier positions drive attenuation. On Qwen7B, noise at earlier positions alone gives clean-pattern coefficients of 0.566 and 0.549 at layers 5 and 14, versus 0.930 and 0.872 for noise at the final position alone. The differences are 0.364 [0.308, 0.423] and 0.323 [0.297, 0.347]. Under the full noise, a falls to 0.548 and 0.452, while the RMS of the mean output stays at 0.967 and 0.957 of its clean value. The noise therefore shrinks the differences between inputs much more than it changes the average output.

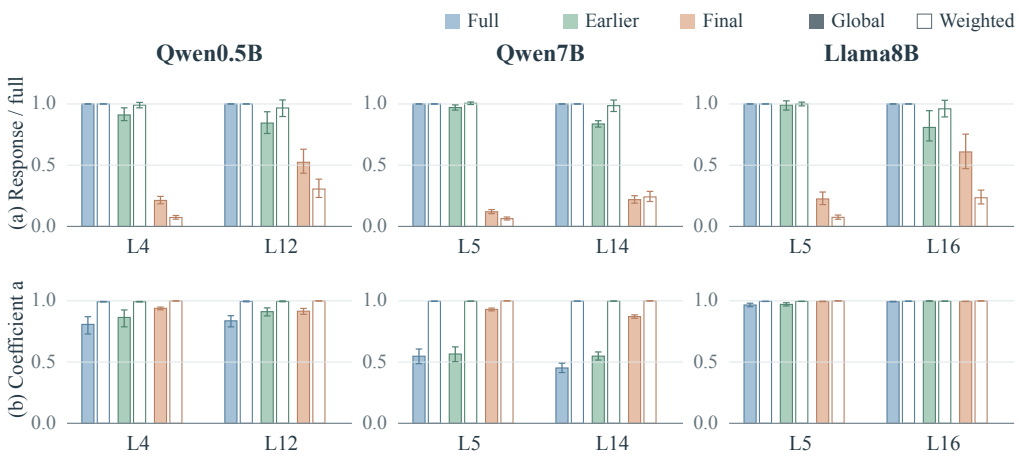

Figure 3: Position and allocation together determine the response. Top: retention, the mean absolute margin change relative to the full noise. Bottom: clean-pattern coefficient a; values below one mean attenuation. Colors mark which positions receive noise; filled and outlined bars mark global and activation-weighted noise. Whiskers give 95% input-bootstrap intervals. Qwen7B uses 80 inputs, the other models 40 each. Noise at earlier positions retains much of the response under both allocations, but strong attenuation appears mainly under global noise.

A large response need not come with attenuation. In Llama, noise at earlier positions retains 99.0% and 80.8% of the full response, yet its clean-pattern coefficients are 0.971 and 0.998. The corresponding Qwen7B coefficients are 0.566 and 0.549. Noise at earlier positions can therefore produce much of the response whether or not it strongly attenuates the clean pattern. Masking positions shows where a response comes from and a shows whether it shrinks the differences between inputs, so together they separate a score change carried from earlier positions from attenuation.

Qwen0.5B shows that attenuation differs from a smaller spread across inputs. The residual $e(x)$ can add new variation while the part aligned with $u(x)$ shrinks. In Qwen7B under full noise, the RMS spread across inputs falls to 0.834 and 0.703 of its clean value, while a is 0.548 and 0.452. At Qwen0.5B layer 4, the spread even increases to 1.13 while a falls to 0.81 (Figure 1b). Attenuation therefore refers only to the shrinking of the clean pattern.

Allocation changes attenuation at fixed total energy. Activation-weighted noise links these position results to the allocation comparison in Section 3.1. At the same two Qwen7B layers, full activation-weighted noise gives clean-pattern coefficients of 0.997 and 0.998, compared with 0.548 and 0.452 under global noise. Its part at earlier positions alone also gives 0.997 and 0.998. Earlier positions still produce almost all of the weighted response but leave the clean pattern intact. With total energy fixed, spreading the noise differently over the same earlier states thus gives very different attenuation (complete comparisons in Appendix B).

5.2 The output model predicts held-out responses

We test whether the output model in Equation 3 predicts responses to new inputs and new noise directions. For each model, we split 160 inputs into five folds. We estimate a and the perturbed mean output μ_D on four folds under one noise direction, predict the held-out fold under the other direction, and then swap the two directions. The coefficient a is fitted to the non-class logits, and the intercept of the predicted margin uses the mean class logits of the training folds. No responses from the held-out fold enter the fit.

Table 1 compares four predictors. Uniform multiplies all centered logits by one fitted factor. Offset lets the mean output change but fixes a at one. Output, our model, also estimates a from the non-class logits. Margin fits a scale and an intercept directly to the class margin. Comparing Output with Margin shows how well a change estimated from the rest of the output carries over to the safety

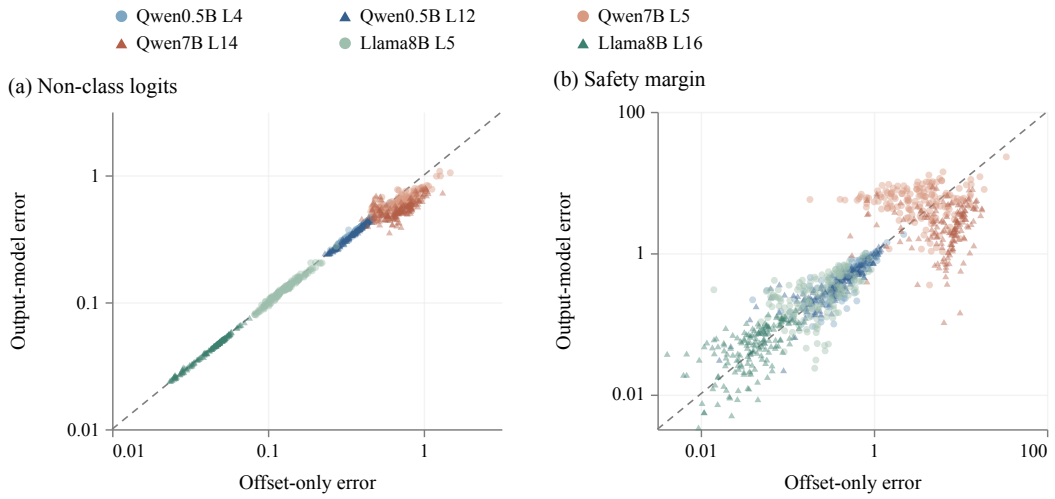

Figure 4: Per-input prediction errors across three models and six layers. Each panel contains 960 points, with 160 inputs per model and layer. A point is the root mean squared error over the two held-out noise directions. Points below the diagonal are inputs for which estimating a (Output) improves on a mean shift alone (Offset). Both axes are logarithmic.

decision. We report the signed margins as well as their mean, since averaging over the two signs can hide variation between them.

At Qwen7B layer 14, the Output model reduces the paired-mean margin RMSE from 7.164 under uniform scaling and 8.600 under a mean shift to 3.315, close to 3.279 for the direct margin fit. The same ordering holds for individual signed responses, with 4.254 versus 7.644, 9.004, and 4.226, respectively.

Per-input errors. Figure 4 compares the errors for each held-out input rather than only their average. At Qwen7B layer 14, compared with Offset, the Output model better predicts the non-class logits for 142 of 160 inputs and the margin for 141 of 160 inputs. The two targets can disagree, however. At layer 5, the non-class RMSE decreases from 0.744 to 0.635, while the margin RMSE increases from 6.167 to 6.580. A better fit to the rest of the output does not always improve the margin prediction by the same amount.

Table 1: Held-out margin RMSE. Each cell reports the paired mean, the average margin over the two noise signs (top), and the individual signed responses (bottom). Lower is better; bold marks the best value in each row for each target.

Model / layer	Uniform	Offset	Output	Margin
Qwen0.5B / 4	0.816	0.583	**0.546**	**0.546**
	1.139	0.986	**0.965**	**0.965**
Qwen0.5B / 12	0.582	0.543	**0.502**	**0.502**
	1.036	1.015	**0.994**	**0.994**
Qwen7B / 5	6.254	6.167	6.580	**5.626**
	7.863	7.795	8.125	**7.374**
Qwen7B / 14	7.164	8.600	3.315	**3.279**
	7.644	9.004	4.254	**4.226**
Llama8B / 5	0.346	0.338	0.348	**0.328**
	0.805	0.801	0.805	**0.797**
Llama8B / 16	0.059	0.056	0.064	**0.055**
	0.317	**0.317**	0.318	**0.317**

Both the mean and the input differences are needed. At Qwen7B layer 14, the non-class RMSE is 0.552 for the full Output model. It is 0.877 for uniform scaling, 0.726 for a mean shift alone, which keeps the clean pattern at full size ($a = 1$), and 0.690 for the shared mean alone, which drops the clean pattern ($a = 0$). Estimating the two changes separately predicts the broad output change and, at this layer, most of the predictable change in the safety margin. Table 1 and Figure 4 contrast this strong attenuation with settings where the clean pattern is largely kept.

Paired and individual responses. In Table 1, individual signed responses have larger errors than paired means. Averaging over the two signs cancels the sign-dependent part of the response, so comparing the two targets separates a predictable change shared by both signs from the variation between opposite perturbations.

5.3 Layer profiles combine a shared score change with layer differences

A separate analysis asks how much of the response profile comes from a score change shared across inputs. For each held-out block of inputs, we fit an affine map from the perturbed paired-mean margin back to the clean margin, using the other blocks and the other noise direction. Unlike the output model above, which predicts perturbed outputs from clean ones, this inverse map predicts clean margins, and it covers every layer instead of only the selected sites.

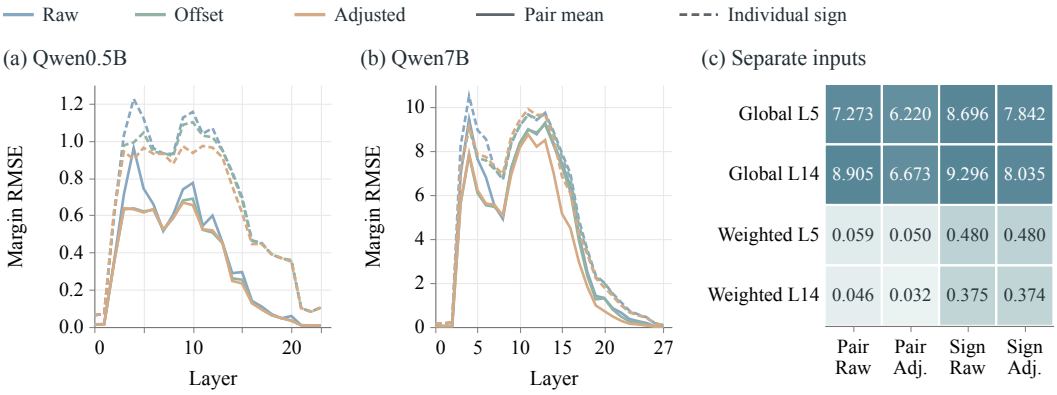

Figure 5: Shared score changes and layer differences. (a,b) RMSE between perturbed and clean margins at every layer, for raw scores and for offset and affine corrections. Solid lines show paired means; dashed lines show individual signs. Maps use other blocks and the other noise sample. (c) Raw and adjusted RMSE for paired-mean and signed margins under both allocations on 160 separate Qwen7B inputs, with maps fitted on 80 other inputs. Cell shading follows logarithmic RMSE.

Across all 24 Qwen0.5B and 28 Qwen7B layers, the correction lowers the mean squared difference between paired-mean and clean margins by 24.0% and 22.8%. Yet the layer ranking by this error barely changes, with rank correlations of 0.977 and 0.984 between the raw and corrected profiles. A score change shared across inputs thus explains much of the response size, while the broad differences between layers remain. Maps fitted on 80 Qwen7B inputs also reduce this error on 160 separate inputs (Figure 5c). Appendix A.4 gives all settings, intervals, and signed-response results.

6 Consequences for Safety Evaluation

A flip-rate profile counts changed decisions but does not say whether the changes help or hurt. For labeled binary predictions, let p_c be the fraction of all inputs whose errors are corrected and p_n the fraction whose correct predictions become errors. Then

$$\text{FlipRate} = p_c + p_n, \qquad \Delta\text{Accuracy} = p_c - p_n. \tag{5}$$

The flip rate adds the two, whereas the change in accuracy depends on their difference. The experiment with opposite noise signs gives a direct example. Of the 25 global pairs in which both signs change the decision after the gradient component is removed, seven correct an error and 18 introduce one. Among these flips under both signs, new errors outnumber corrections. The larger model-size comparison in Table 17 gives more examples, and Table 10 splits the flips of the component comparisons into corrections and new errors.

Moving scores back toward their clean values does not necessarily improve classification either. When the inverse map described above is applied to new inputs, the Qwen7B layer-5 RMSE decreases while accuracy falls from 75.94% to 66.88%. This map targets clean margins (Section 5.2 predicts perturbed ones), whereas accuracy depends on agreement with the dataset labels. Component comparisons within each model and all inverse-map results appear in Appendices C and A.4.

The separate prompt-formatting experiments in Appendix L follow the same principle and judge changes in harmful responses together with changes in ordinary answering. For Qwen0.5B, the lower attack success under prompt formatting comes with a large loss of ordinary answering.

7 DISCUSSION

The intervention determines what is measured. Two allocations with the same energy differ in how much their responses depend on the clean decision direction. Position masks show how noise at earlier states reaches the output. Together they separate where a response comes from and what it does. In Qwen7B, noise at earlier states produces large responses under both allocations, but weighted noise largely keeps the clean pattern while global noise strongly attenuates it. A layer's response size is most useful when reported together with the noise components and output changes behind it.

Input differences under a shared output change. The mean output and each input's deviation from it change by different amounts. Keeping the deviations improves held-out prediction over the mean alone, and fitting their change separately improves on uniform scaling, although residual variation can remain or grow. The fitted model thus predicts from each input's clean deviation how its response differs from those of other inputs.

Predictable score shifts and accuracy. For $a > 0$, an affine prediction $\hat{m}_D(x) = a\, m_0(x) + b$ keeps the order of the margins but moves the decision threshold to $-b/a$. Corrections and new errors then depend on which labeled examples cross that threshold. The same predictable response can therefore help in one labeled setting and hurt in another. Position tests and prediction across inputs describe the response, but only labeled harmful and benign outcomes show whether it helps. Open questions include which computations in later layers cause the attenuation and whether it predicts refusal in free-form generation.

8 RELATED WORK

Safety measurement and latent interventions. HELM, DecodingTrust, and HarmBench organize behavioral evaluation (Liang et al., 2022; Wang et al., 2023; Mazeika et al., 2024); OR-Bench and XSTest measure over-refusal (Cui et al., 2024; Röttger et al., 2024). Activation engineering and representation engineering control outputs through hidden states (Turner et al., 2023; Zou et al., 2023a; Panickssery et al., 2024). ASA optimizes latent attacks (Gu et al., 2025), and Safety Layers and LayerNavigator identify layers for intervention (Li et al., 2025; Sun et al., 2025). We study why the measured responses arise and test whether selected layers stay responsive on later inputs.

Mechanistic explanations. Studies of refusal describe it by a single direction (Arditi et al., 2024) or by multidimensional structures (Wollschläger et al., 2025), and separate harmfulness from refusal (Zhao et al., 2025). Safety can depend on the first generated tokens (Qi et al., 2024), whereas our position interventions act on prompt states before generation. Activation patching traces internal pathways (Meng et al., 2022; Zhang & Nanda, 2024; Cheng et al., 2026), and random steering directions can cause shifts that learned directions also cause (Malla et al., 2026). We relate the noise allocation to how much the response depends on the decision gradient and to the positions that pass it on, and we predict safety margins on new inputs and noise from changes in non-class outputs.

Predicting output changes. The tuned lens decodes intermediate states into vocabulary distributions (Belrose et al., 2023), and control tasks separate what a representation contains from what a probe learns (Hewitt & Liang, 2019). We instead model how a perturbation changes the final outputs and use the changes in non-class outputs to predict safety margins. Appendix O discusses these and other related studies in more detail.

9 CONCLUSION

Noise with the same energy can change safety judgments in different ways. Its allocation determines which noise directions produce the response, and position interventions show where the response enters. Changes in non-class outputs predict decision margins on new inputs and noise, and labeled outcomes show whether these changes improve accuracy. Together, these measurements link a response peak to the applied noise, the affected positions, and the resulting output changes.

AI USE STATEMENT

AI tools assisted with drafting and polishing the manuscript and with retrieving and discovering related literature. The authors reviewed and verified all resulting text and references and take responsibility for the final manuscript.

ETHICS STATEMENT

The study examines safety judgments and refusal behavior using existing benchmarks. Such data can contain harmful language, and sensitivity measurements could also inform attempts to bypass safeguards. We report aggregate results and avoid reproducing actionable harmful prompts. Model and dataset terms are summarized in Appendix N; releases should preserve source licensing and avoid disclosing private conversation content.

REPRODUCIBILITY STATEMENT

Section 2 defines the readouts, models, and comparisons. Appendices A and B specify the direction, position, and analyses of output prediction. Appendices D, E, and H document the original perturbations and uncertainty estimates; Appendix I gives the rubric for human review. Comparisons of layer selection budgets and model/data sources appear in Appendices K and N.

REFERENCES

Maksym Andriushchenko and Nicolas Flammarion. Does refusal training in llms generalize to the past tense? *arXiv preprint arXiv:2407.11969*, 2024.

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.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*, 2022.

Nora Belrose, Igor Ostrovsky, Lev McKinney, Zach Furman, Logan Smith, Danny Halawi, Stella Biderman, and Jacob Steinhardt. Eliciting Latent Predictions from Transformers with the Tuned Lens. *arXiv preprint arXiv:2303.08112*, 2023. URL https://arxiv.org/abs/2303.08112.

Collin Burns, Haotian Ye, Dan Klein, and Jacob Steinhardt. Discovering Latent Knowledge in Language Models Without Supervision. *arXiv preprint arXiv:2212.03827*, 2022. URL https://arxiv.org/abs/2212.03827.

Patrick Chao, Alexander Robey, Edgar Dobriban, Hamed Hassani, George J. Pappas, and Eric Wong. Jailbreaking Black Box Large Language Models in Twenty Queries. *arXiv preprint arXiv:2310.08419*, 2023. URL https://arxiv.org/abs/2310.08419.

Cléa Chataigner, Rebecca Ma, Prakhar Ganesh, Yuhao Chen, Afaf Taïk, Elliot Creager, and Golnoosh Farnadi. Say it another way: Auditing llms with a user-grounded automated paraphrasing framework. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1441–1467, 2026.

Stephen Cheng, Sarah Wiegreffe, and Dinesh Manocha. What drives representation steering? a mechanistic case study on steering refusal, 2026. URL https://arxiv.org/abs/2604.08524.

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.

Deep Ganguli, Liane Lovitt, Jackson Kernion, Amanda Askell, Yuntao Bai, Saurav Kadavath, Ben Mann, Ethan Perez, Nicholas Schiefer, Kamal Ndousse, et al. Red teaming language models to reduce harms: Methods, scaling behaviors, and lessons learned. *arXiv preprint arXiv:2209.07858*, 2022.

Atticus Geiger, Hanson Lu, Thomas Icard, and Christopher Potts. Causal Abstractions of Neural Networks. *arXiv preprint arXiv:2106.02997*, 2021. URL https://arxiv.org/abs/2106.02997.

Nicholas Goldowsky-Dill, Chris MacLeod, Lucas Sato, and Aryaman Arora. Localizing Model Behavior with Path Patching. *arXiv preprint arXiv:2304.05969*, 2023. URL https://arxiv.org/abs/2304.05969.

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. URL https://arxiv.org/abs/2407.21783.

Tianle Gu, Kexin Huang, Zongqi Wang, Yixu Wang, Jie Li, Yuanqi Yao, Yang Yao, Yujiu Yang, Yan Teng, and Yingchun Wang. Probing the robustness of large language models safety to latent perturbations, 2025. URL https://arxiv.org/abs/2506.16078.

John Hewitt and Percy Liang. Designing and Interpreting Probes with Control Tasks. *arXiv preprint arXiv:1909.03368*, 2019. URL https://arxiv.org/abs/1909.03368.

Xiaomeng Hu, Pin-Yu Chen, and Tsung-Yi Ho. Gradient cuff: Detecting jailbreak attacks on large language models by exploring refusal loss landscapes. *Advances in Neural Information Processing Systems*, 37:126265–126296, 2024.

Andong Hua, Kenan Tang, Chenhe Gu, Jindong Gu, Eric Wong, and Yao Qin. Flaw or artifact? rethinking prompt sensitivity in evaluating llms. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 19889–19899, 2025.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. Mistral 7B. *arXiv preprint arXiv:2310.06825*, 2023. URL https://arxiv.org/abs/2310.06825.

Shen Li, Liuyi Yao, Lan Zhang, and Yaliang Li. Safety layers in aligned large language models: The key to llm security, 2025. URL https://arxiv.org/abs/2408.17003.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, et al. Holistic evaluation of language models. *arXiv preprint arXiv:2211.09110*, 2022.

Zi Lin, Zihan Wang, Yongqi Tong, Yangkun Wang, Yuxin Guo, Yujia Wang, and Jingbo Shang. ToxicChat: Unveiling Hidden Challenges of Toxicity Detection in Real-World User-AI Conversation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 4694–4702, 2023. doi: 10.18653/v1/2023.findings-emnlp.311. URL https://aclanthology.org/2023.findings-emnlp.311/.

Xiaoya Lu, Dongrui Liu, Yi Yu, Luxin Xu, and Jing Shao. X-boundary: Establishing exact safety boundary to shield llms from jailbreak attacks without compromising usability. *Findings of the Association for Computational Linguistics: EMNLP*, 2025:5247–5272, 2025.

Srikanth Malla, Chiho Choi, and Joon Hee Choi. Steering interference reflects the model's defaults, not the behavior directions. *arXiv preprint arXiv:2609.06951*, 2026. URL https://arxiv.org/abs/2609.06951.

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.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and Editing Factual Associations in GPT. *arXiv preprint arXiv:2202.05262*, 2022. URL https://arxiv.org/abs/2202.05262.

Licheng Pan, Yongqi Tong, Xin Zhang, Xiaolu Zhang, Jun Zhou, and Zhixuan Chu. Understanding and mitigating overrefusal in llms from an unveiling perspective of safety decision boundary. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 21057–21075, 2025.

Nina Panickssery, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Matt Turner. Steering Llama 2 via contrastive activation addition. *arXiv preprint arXiv:2312.06681*, 2024. URL https://arxiv.org/abs/2312.06681.

Ethan Perez, Saffron Huang, Francis Song, Trevor Cai, Roman Ring, John Aslanides, Amelia Glaese, Nat McAleese, and Geoffrey Irving. Red Teaming Language Models with Language Models. *arXiv preprint arXiv:2202.03286*, 2022. URL https://arxiv.org/abs/2202.03286.

Xiangyu Qi, Ashwinee Panda, Kaifeng Lyu, Xiao Ma, Subhrajit Roy, Ahmad Beirami, Prateek Mittal, and Peter Henderson. Safety alignment should be made more than just a few tokens deep, 2024. URL https://arxiv.org/abs/2406.05946.

Qwen, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, et al. Qwen2.5 Technical Report. *arXiv preprint arXiv:2412.15115*, 2024. URL https://arxiv.org/abs/2412.15115.

Paulius Rauba, Qiyao Wei, and Mihaela Van Der Schaar. Statistical hypothesis testing for auditing robustness in language models. *arXiv preprint arXiv:2506.07947*, 2025.

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.

Hao Sun, Huailiang Peng, Qiong Dai, Xu Bai, and Yanan Cao. Layernavigator: Finding promising intervention layers for efficient activation steering in large language models. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 101058–101080. Curran Associates, Inc., 2025. doi: 10.52202/085713-3383. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/9280bfb61f159c768dfdeb729a41ccd0-Paper-Conference.pdf.

Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J Vazquez, Ulisse Mini, and Monte MacDiarmid. Steering language models with activation engineering. *arXiv preprint arXiv:2308.10248*, 2023.

Boxin Wang, Weixin Chen, Hengzhi Pei, Chulin Xie, Mintong Kang, Chenhui Zhang, Chejian Xu, Zidi Xiong, Ritik Dutta, Rylan Schaeffer, et al. Decodingtrust: A comprehensive assessment of trustworthiness in GPT models. In *Advances in Neural Information Processing Systems*, volume 36, pp. 31232–31339, 2023.

Alexander Wei, Nika Haghtalab, and Jacob Steinhardt. Jailbroken: How does llm safety training fail? *Advances in neural information processing systems*, 36:80079–80110, 2023.

Tom Wollschläger, Jannes Elstner, Simon Geisler, Vincent Cohen-Addad, Stephan Günnemann, and Johannes Gasteiger. The geometry of refusal in large language models: Concept cones and representational independence. *arXiv preprint arXiv:2502.17420*, 2025.

Chen Xiong, Zhiyuan He, Pin-Yu Chen, Ching-Yun Ko, and Tsung-Yi Ho. Steering externalities: Benign activation steering unintentionally increases jailbreak risk for large language models. *arXiv preprint arXiv:2602.04896*, 2026.

Zhao Xu, Fan Liu, and Hao Liu. Bag of tricks: Benchmarking of jailbreak attacks on llms. *Advances in Neural Information Processing Systems*, 37:32219–32250, 2024.

Fred Zhang and Neel Nanda. Towards Best Practices of Activation Patching in Language Models: Metrics and Methods. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=Hf17y6u9BC.

Jiachen Zhao, Jing Huang, Zhengxuan Wu, David Bau, and Weiyan Shi. Llms encode harmfulness and refusal separately. *arXiv preprint arXiv:2507.11878*, 2025.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36:46595–46623, 2023.

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*, 2023a.

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.

A Interpreting Responses to Latent Perturbations

A.1 Decision score, perturbations, and evaluation units

These experiments study the next-token toxicity decision under each model's official chat template. They use Qwen2.5-0.5B-Instruct, Qwen2.5-7B-Instruct, and Llama-3.1-8B-Instruct to measure continuous changes in class scores. The system message instructs the model to classify quoted text, not follow it. The user asks whether the text is toxic, requesting only `Yes` or `No`. The input follows as a JSON string. The margin m is the logit of the positive class token minus that of the negative class token; its sign defines the binary decision. We retain every margin and separately record whether the unrestricted top-ranked token is a class answer.

For a clean layer output $H = (h_1, \ldots, h_T)$, let Z be an independent standard Gaussian tensor of the same shape. The globally scaled perturbation is $D_G = \delta \,\mathrm{sd}(H) Z$. Let $r_t = \sqrt{d^{-1} \sum_j h_{tj}^2}$ denote a token's activation RMS. The activation-weighted perturbation is

$$D_R = \|D_G\|_F \frac{r \odot D_G}{\|r \odot D_G\|_F}. \tag{6}$$

Thus both operators have the same total squared norm before casting to the model's precision, while allocating it differently across tokens. A control with shuffled weights permutes r across positions before the same normalization. The direction experiments and the experiments that transfer a fitted score correction to new inputs (Appendix A.4) act on the full block output at all prompt tokens; layers are numbered from zero. Opposite signs reuse the same sampled tensor. The strength is $\delta = 0.025$ throughout; layers and magnitude are chosen before evaluating the responses.

The development set contains 80 examples, balanced by toxicity label within the five data blocks. The pilot removing the gradient component uses 20 of these inputs, with four layers at relative depths 0, approximately $1/6$, $1/2$, and the last layer. The gradient experiment on additional inputs uses the other 60 Qwen7B inputs and layers 5 and 14. The analysis across all layers uses the 80 development inputs with their original partition into five blocks. The final transfer experiment fits the score correction on these 80 inputs and applies it to 160 non-overlapping inputs. Those 160 inputs had appeared in a separate generation evaluation, but were not used to fit or select the score corrections. Uncertainty is estimated across inputs, retaining their layers, directions, and signs together.

A.2 Dependence on the clean decision direction

For each token, compute $g_t = \nabla_{h_t} m(H)$ and replace a sampled perturbation d_t by its component orthogonal to g_t, then rescale it to restore $\|d_t\|_2$. Tokens with gradient norm at most 10^{-20} are left unchanged and counted. This operation preserves the per-token perturbation norm, and hence the allocation and total energy, while removing the component along the clean score gradient. The gradient experiments use FP32 inference with TF32 disabled; projection and norm calculations use float64 before casting. In the additional 60-input evaluation, the clean run and a replay without noise give identical margins, and the maximum per-token norm error before casting is below 3.4×10^{-9}.

Table 2: Response retained after removing the clean component along the decision gradient (%). Allocations share total energy before projection; projection preserves each token's perturbation norm. The 20-input comparisons use four layers per model; the additional 60-input Qwen7B comparison uses layers 5 and 14. Brackets give 95% bootstrap intervals over inputs for the latter.

Model	Inputs	Global	Activation-weighted	Shuffled weights
Qwen0.5B	20	71.62	16.65	70.32
Qwen7B	20	97.25	10.77	85.32
Llama8B	20	39.14	8.73	70.23
Qwen7B	60	96.01 [93.91,98.04]	9.99 [8.49,11.97]	not evaluated

Table 2 shows that allocation strongly changes the dependence on the clean decision direction. In the 60-input experiment, mean absolute response is 7.246 before and 6.956 after projection for global

noise, compared with 0.365 and 0.037 for activation-weighted noise. Projection removes linear and nonlinear contributions along the clean gradient. Activation weighting therefore concentrates the response along that direction, whereas global allocation also produces large responses in orthogonal directions.

Let G_ℓ and R_ℓ be the retained responses under global and activation-weighted noise at layer ℓ. The layer contrast, $(G_{14} - G_5) - (R_{14} - R_5)$ after projection, is 0.687 with a 95% interval of $[-0.605, 1.917]$ on the additional inputs. Its interval spans zero, so the layer interaction remains uncertain despite the large overall difference in retained response between allocations. Intervals here use 10,000 paired bootstrap samples drawn by input; the 20-input pilot is not pooled into the 60-input result.

A.3 Opposite signs and the shared shift

For paired margins, write

$$E = \frac{m_+ + m_-}{2} - m_0, \qquad O = \frac{m_+ - m_-}{2}, \qquad m_\pm - m_0 = E \pm O. \tag{7}$$

The even term E, which is unchanged when the noise sign is reversed, is the shared shift; the odd term O, which reverses with the sign, is the sign-dependent change. If $|E| > |O|$, both perturbations move the margin in the same direction. If the margin changed affinely along the line from $-D$ to $+D$, the two signs could not both cross the clean decision boundary.

Among 240 pairs under global noise in the 60-input experiment, 24 pairs flip both signs, 39 flip one, and 177 flip neither. After gradient removal these counts are 25, 28, and 187. Every output in this comparison has a valid class top token. Flips under both signs include corrections and new errors: after projection, 7 pairs correct the clean decision and 18 introduce errors. Activation-weighted noise yields no flips under both signs, 9 flips under one sign before projection, and none after projection.

A separate 20-input check changes the answer encoding from A=toxic/B=non-toxic to B=toxic/A=non-toxic, orienting each margin by semantic class. The sign of the shared shift agrees for 15 of 20 inputs at layer 5 and all 20 at layer 14; rank correlations are 0.795 and 0.961. All class outputs are valid. The encodings share seeds and tensor shapes; each uses its own hidden state scale to set the perturbation magnitude. The shift thus follows the meaning of the class rather than a fixed preference for an answer letter.

A.4 A score change shared across inputs and noise

Let $p = (m_+ + m_-)/2$. We fit the ordinary least-squares map $\hat{m} = ap + b$ from the paired-mean margin to the clean margin on fitting inputs only; this slope a is not the clean-pattern coefficient of the main text. A map fitted using one noise realization is evaluated on the other; the reverse assignment is also evaluated. Offset-only, identity, and fits using the same noise provide controls. Toxicity labels are not used in these fits. The affine fit checks whether one score correction, reused across inputs and noise, explains the responses.

Reuse across all layers. For each of the five blocks, the other 64 development inputs fit the map and the held-out 16 evaluate it, with the two noise realizations exchanged. Across all 24 Qwen0.5B layers, MSE of the paired mean falls by 23.99% (95% interval: 12.97–30.98%); across all 28 Qwen7B layers it falls by 22.83% (11.73–29.70%). Yet raw and adjusted layer rankings by RMSE correlate at 0.977 and 0.984 (Figure 5). Individual peak layers can move, while the broad pattern across depth remains. Intervals use 2,000 bootstrap samples drawn by input stratified by block and label, refitting within each sample. This analysis across all layers is exploratory. For individual signed responses pooled over Qwen7B layers, the MSE reduction is smaller, 8.53%, with interval $[-0.26, 13.65]\%$.

Transfer to separate inputs. We estimate maps on all 80 source inputs using two fresh noise realizations, then fix the coefficients and apply each map to the opposite realization on 160 separate inputs. The two layer-14 global slopes are 1.687 and 1.667, versus 1.029 and 0.980 at layer 5. Thus layer 14 benefits from a scale adjustment beyond an offset, whereas layer 5 is mainly an offset change. Table 3 reports every tested allocation and layer. The evaluation contains 2,880 records,

including 160 clean and 160 exact repeats without noise. One perturbed top token is not a class answer; its margin is retained. The total energy of the two allocations matches before casting up to a maximum relative error of 2.1×10^{-7}.

At layer 14, RMSE of the paired mean decreases by 2.232 (95% interval: 1.864–2.611), disagreement falls by 4.69 percentage points (1.88–7.81), and accuracy rises by 5.31 points (2.50–8.44). At layer 5, RMSE also decreases but accuracy falls by 9.06 points (6.25–12.19). Clean accuracy is 71.88%. A separate threshold selected using source labels reaches 87.50% on clean target margins, versus 86.25% on raw and 84.06% on adjusted layer-14 paired means. Restoring clean margins and predicting toxicity labels are different targets, which explains these differences. Primary intervals use 2,000 label-stratified bootstrap samples drawn by input, holding the maps fitted on the source inputs fixed.

Individual signed responses. Applying the same map to each signed margin gives the exact identity

$$\frac{(am_+ + b - m_0)^2 + (am_- + b - m_0)^2}{2} = (ap + b - m_0)^2 + a^2 O^2. \tag{8}$$

A slope above one can improve the paired mean while amplifying the sign-dependent part O. At layer 14, the error of the paired mean falls from 79.305 to 44.529, while the odd term rises from 7.109 to 20.033. Total signed RMSE nevertheless falls from 9.296 to 8.035, a reduction of 1.261 (95% interval: 0.894–1.612). Signed disagreement falls from 18.91% to 15.00%. This secondary analysis uses the same fixed coefficients after the primary evaluation of the paired mean, with no new inference or target fitting. A mean correction that carries over to new inputs can therefore coexist with substantial remaining sign-dependent variation.

B Position Interventions and Prediction of Output Changes

B.1 Inputs, pairing, and decision score

The position and vocabulary analyses use the toxicity margin with the official chat template defined in Appendix A. Qwen2.5-0.5B-Instruct and Qwen2.5-7B-Instruct use their official pretrained instruction checkpoints; the third model is Llama-3.1-8B-Instruct. All interventions change a block output at magnitude 0.025, with two independently sampled Gaussian directions and their negatives. The preselected layers are 4 and 12, 5 and 14, and 5 and 16, respectively. For every model and input, the evaluation includes a clean output and a repeat without noise. The continuous margin is retained even when the unrestricted next-token argmax is not a class answer.

The Qwen7B forward-prediction analysis, which predicts perturbed outputs from clean ones, reuses the 160 inputs from the score-correction transfer experiment (Appendix A.4), whose outcomes had already been examined. We therefore report its five-fold prediction as an exploratory reanalysis with held-out fitting. The same specified comparisons were subsequently evaluated on Qwen0.5B and Llama8B using 160 inputs each. The saved vocabulary logits cover clean outputs and both signs of two directions under both allocations at both layers.

Table 3: Transfer of the fitted score correction to 160 separate Qwen7B inputs. RMSE is relative to the clean margin. Disagreement and accuracy use the sign of the paired mean margin; accuracy uses toxicity labels only for evaluation. Adjusted values use the fixed affine map fitted under the other noise direction. At layer 5, closer agreement with clean margins accompanies lower classification accuracy.

		Paired mean RMSE			Disagreement (%)		Accuracy (%)	
Allocation	Layer	Raw	Offset	Adjusted	Raw	Adjusted	Raw	Adjusted
Global	5	7.27	6.22	6.22	12.19	14.38	75.94	66.88
Global	14	8.91	8.57	6.67	19.69	15.00	53.44	58.75
Weighted	5	0.059	0.055	0.050	0.31	0.63	71.56	71.88
Weighted	14	0.046	0.046	0.032	0.63	0.63	71.25	71.25

For position decomposition, the initial Qwen7B analysis uses the first 40 inputs ordered by hash from this group. The subsequent position comparison uses the next 80, with no overlap in outcomes of position interventions. Their clean outputs and outputs under full noise had previously been observed. Qwen0.5B and Llama8B use the first 40 corresponding hashes. The Qwen7B initial and subsequent position groups each contain one non-class top-token output; both are retained. The two other models have none in these comparisons. The input groups and their results are reported separately.

B.2 Complementary position interventions

Let P_E select every prompt position before the last and let P_F select only the last prompt position. Given the original tensor D, define $D_E = P_E \odot D$ and $D_F = P_F \odot D$. Their sum equals D before and after casting to inference precision. Neither component is renormalized. The last position includes the official chat template suffix and is the position from which the next token is predicted. It is the template position used to read out the answer. Only the intervention layer is masked; later layers can pass changes between positions.

For each noise part $s \in \{\text{full}, E, F\}$, define $\Delta m_s = m(H + D_s) - m(H)$. Response retention is the ratio of mean absolute changes,

$$R_s = \frac{\mathbb{E}_{x,D}|\Delta m_s|}{\mathbb{E}_{x,D}|\Delta m_{\text{full}}|}.$$

The signed interaction, $I = \Delta m_{\text{full}} - \Delta m_E - \Delta m_F$, is the part of the full change that the sum of the two parts does not explain. Each expectation averages the same inputs, directions, and signs. Each ratio measures a response relative to the full intervention; their sum can differ from one because of nonlinear interaction and the absolute value. Earlier positions contain about 98.9% of original energy under global noise across these groups, whereas the activation-weighted energy fraction can be still larger. The masks keep these energy fractions, so each part's contribution is measured under the original intervention.

Table 4: Global allocation: responses and output coefficients for each position group. Retention is a ratio of mean absolute responses. Brackets are 95% bootstrap intervals over inputs. The regression coefficient a measures the scale of the clean output pattern.

Model / layer	Earlier retention	Final retention	Full a	Earlier a	Final a
Qwen0.5B / 4	91.0	21.2	0.808	0.864	0.939
	[86.4,96.8]	[18.4,24.5]	[0.729,0.871]	[0.788,0.926]	[0.927,0.949]
Qwen0.5B / 12	84.4	52.4	0.837	0.912	0.916
	[75.9,93.6]	[43.4,63.0]	[0.787,0.879]	[0.877,0.943]	[0.890,0.938]
Qwen7B / 5	97.1	12.0	0.548	0.566	0.930
	[95.0,99.2]	[10.4,13.7]	[0.487,0.607]	[0.504,0.624]	[0.918,0.941]
Qwen7B / 14	83.7	21.8	0.452	0.549	0.872
	[81.1,86.3]	[19.0,25.1]	[0.414,0.491]	[0.516,0.583]	[0.858,0.886]
Llama8B / 5	99.0	22.4	0.966	0.971	0.995
	[95.0,102.6]	[17.8,28.1]	[0.952,0.981]	[0.958,0.985]	[0.994,0.997]
Llama8B / 16	80.8	60.8	0.994	0.998	0.996
	[69.8,94.5]	[47.2,75.4]	[0.992,0.996]	[0.996,1.001]	[0.995,0.997]

The comparisons using global noise retain large responses from earlier positions across all six sites. Responses from the final position become more substantial at the deeper Qwen0.5B and Llama sites, while Qwen7B shows the clearest separation in output coefficients.

The weighted controls largely preserve the output pattern associated with clean inputs despite the same total perturbation energy as global allocation. The position split preserves the original energy distribution within each allocation.

Table 5: Activation-weighted allocation: responses and output coefficients for each position group. Retention is a ratio of mean absolute responses. Brackets are 95% bootstrap intervals over inputs. The regression coefficient a measures the scale of the clean output pattern.

Model / layer	Earlier retention	Final retention	Full a	Earlier a	Final a
Qwen0.5B / 4	99.1	7.3	0.993	0.993	1.000
	[96.8,101.3]	[5.9,8.9]	[0.991,0.995]	[0.991,0.995]	[0.999,1.000]
Qwen0.5B / 12	96.7	30.5	0.996	0.996	1.000
	[89.7,103.3]	[23.6,38.6]	[0.992,0.999]	[0.992,1.000]	[0.999,1.000]
Qwen7B / 5	100.6	6.3	0.997	0.997	1.000
	[99.5,101.6]	[5.3,7.6]	[0.997,0.998]	[0.997,0.998]	[1.000,1.000]
Qwen7B / 14	98.6	24.1	0.998	0.998	1.000
	[93.8,103.2]	[20.3,28.6]	[0.997,0.998]	[0.997,0.998]	[1.000,1.000]
Llama8B / 5	100.0	7.4	0.996	0.996	1.000
	[98.4,101.5]	[5.8,9.2]	[0.995,0.997]	[0.995,0.997]	[1.000,1.000]
Llama8B / 16	96.0	23.4	0.997	0.997	1.000
	[89.4,103.0]	[18.3,29.6]	[0.997,0.998]	[0.997,0.998]	[0.999,1.000]

B.3 Vocabulary centering and the fitted component

For vocabulary size V, we remove the arbitrary common logit offset using $z_v(x) = l_v(x) - V^{-1}\sum_w l_w(x)$. We then compute the mean across inputs μ_v separately for clean and perturbed outputs. Special tokens and the coordinates for the class answers are excluded when estimating a, but vocabulary centering precedes that exclusion. Using fitting inputs $\mathcal{T}$ and retained coordinates $\mathcal{V}$, the least-squares coefficient is

$$a = \frac{\sum_{x\in\mathcal{T}}\sum_{v\in\mathcal{V}}[z_v(x)-\mu_v][z_{D,v}(x)-\mu_{D,v}]}{\sum_{x\in\mathcal{T}}\sum_{v\in\mathcal{V}}[z_v(x)-\mu_v]^2}. \tag{9}$$

The coefficient is shared across coordinates and inputs for a specified model, layer, allocation, and fitting direction. A value below one means that the part of the output aligned with the clean pattern, the clean deviations across inputs, shrinks; we call this attenuation. The residual captures other output variation.

The position analysis fits this coefficient on the mean output over two directions and both signs. RMS ratio of the mean output and cosine similarity describe μ_D relative to μ. Total input variation is measured separately by the RMS of $z_D - \mu_D$ relative to $z - \mu$. For Qwen7B full global noise, the ratios of total variation are 0.834 and 0.703 at layers 5 and 14, different from coefficients 0.548 and 0.452. For Qwen0.5B the corresponding ratios exceed one despite coefficients below one. The clean pattern can therefore weaken while residual variation increases the total spread across inputs.

B.4 Forward prediction on held-out inputs and noise

Five deterministic folds are stratified by toxicity label using seed 20260922; labels balance the folds but do not enter the regression. In each fold, coefficients and clean/perturbed vocabulary means are estimated from the other four folds under one noise direction. Predictions are evaluated on the held-out inputs under the other direction, and the direction assignment is reversed. The primary target is the mean of the positive and negative outputs. We also evaluate individual signed outputs from the same records.

The output model predicts $\hat{z}_D(x) = \mu_{D,\mathcal{T}} + a[z(x) - \mu_{\mathcal{T}}]$. Its prediction of the class margin uses the class means from the fitting inputs to construct the intercept; only the gain is fitted outside the class coordinates. Identity, uniform temperature scaling, shared-offset ($a = 1$), shared-only ($a = 0$), and separately fitted affine models of the binary margin are evaluated on exactly the same folds and targets. Uniform scaling uses one multiplicative coefficient on the vocabulary-centered logit vectors, without a separate offset for each coordinate. Shared offset allows a separate mean change per vocabulary coordinate. The affine model of the binary margin fits its scale and intercept directly to the pairs of class margins. This control shows how much is lost by fitting the broader output instead of the class margin directly.

At Qwen7B layer 14 under global noise, non-class vocabulary RMSE is 0.931 for identity, 0.877 for uniform scaling, 0.726 for shared offset, 0.690 for shared-only prediction, and 0.552 for the output model. Prediction also improves on the 256 highest-ranked retained clean-output coordinates. The mean fitted coefficient is about 0.470; the shared mean has RMS ratio about 0.958 and cosine about 0.975. The total RMS ratio across inputs is 0.778, reflecting both the fitted component and residual variation.

Table 6: Output structure and prediction. Non-class vocabulary RMSE at Qwen7B layer 14 on the same held-out combinations of inputs and noise. Each restricted model tests a different explanation; the full output model retains a mean output and a scaled component associated with each input.

Model	Restriction on the output	RMSE
Identity	No output change	0.931
Uniform scaling	One gain for mean and deviations	0.877
Shared offset	Preserve clean deviations	0.726
Shared only	Discard clean deviations	0.690
Output model	Change mean and deviation scale separately	0.552

Table 7: Forward prediction at Qwen7B layer 14. Both rows evaluate the same 160 inputs with held-out folds and exchanged noise directions. The signed target retains direction-dependent variation that the paired mean removes. Errors are decision margin RMSE.

Target	Identity	Uniform scaling	Shared offset	Output model	Margin fit
Paired mean	8.905	7.164	8.600	3.315	3.279
Individual sign	9.296	7.644	9.004	4.254	4.226

Qwen7B layer 5 also shows gains in vocabulary prediction: the output model's non-class RMSE is 0.635, compared with 0.744 for an offset and 0.861 for uniform scaling. However, class margin RMSE is 6.580, worse than the offset's 6.167, uniform scaling's 6.254, and the binary fit's 5.626. A good fit to the vocabulary change does not always explain the class logits equally well. At Qwen7B's two activation-weighted sites, the coefficients remain about 0.997 and errors in predicting the mean margin over opposite signs are 0.051 and 0.032. Under global noise, Qwen0.5B has coefficients around 0.841–0.846 in the 160-input analysis, and Llama8B has coefficients around 0.963–0.992. The complete global comparison of class margins is in Table 1.

B.5 Uncertainty and interpretation

Intervals for the position responses use 2,000 bootstrap samples drawn by input, retaining all signs and directions from each sampled input. Output coefficients and their differences are refitted in each resample using the stored sufficient statistics. Intervals for forward prediction resample the saved held-out predictions by input, conditional on the fitted folds; they do not include the variation of refitting the folds. We did not formally test equivalence between the errors of the output model and the affine margin model.

The position experiments show that noise at earlier states can produce both a margin response and attenuation of the clean pattern, and the prediction experiments show that one output model carries over to other inputs and noise. Which attention and MLP operations implement this transformation remains open. A separate analysis of content restoration found no clear evidence that semantic content is selectively destroyed.

C Judge Phrasing and Component Comparisons

We compare decisions obtained from alternative judge phrasings and measure residual and attention responses at the same layers. Human review of the generated responses is described in Appendix I.

The calibrated phrasings agree on at least 90% of records, and adjudication removes the remaining phrasing disagreements. These values measure how stable the extracted labels are across phrasings; the semantic review is in Appendix I.

Table 8: Alternative judge phrasings yield similar labels. Results use 80 examples per dataset; accuracy compares the majority vote with the dataset label.

Setting / Dataset	Min. pair agreement	Spread	κ	Majority accuracy	Majority positive
Calibrated / JBB	0.9000	0.0625	−0.0345	0.5000	0.5000
Calibrated / XSTest	0.9750	0.0250	−0.0084	0.5750	0.4250
Adjudicated / JBB	1.0000	0.0000	1.0000	0.5125	0.4750
Adjudicated / XSTest	1.0000	0.0000	1.0000	0.5625	0.4250

C.1 Component Responses and Accuracy

Table 9 compares flip rates, accuracy, and flip directions for residual and attention interventions. Grouping settings by observed Kullback–Leibler (KL) divergence of the decision logits makes their output changes more comparable.

These comparisons use 500 inputs, 241 toxic and 259 non-toxic, so always predicting non-toxic gives 51.8% accuracy. Clean accuracy is 63.6% for Qwen7B and 67.4% for Mistral7B. Changes in accuracy below are measured from each model's clean predictions on these inputs.

Table 9: Component comparisons grouped by observed KL divergence of the decision logits. R/A denote residual/attention; the gap is R minus A flip rate. Accuracy cells give the accuracy and, below it, the change from clean in percentage points. H/B denote harmful/benign predictions. Values are averages over four KL target settings. A is the attention branch with the larger mean flip rate (the output branch in case of a tie).

Layer	R flip	A flip	Gap	R accuracy (Δ pp)	A accuracy (Δ pp)	H to B	B to H
Qwen7B							
13	0.132	0.031	+0.101	0.516 (−12.00)	0.620 (−1.60)	0.122	0.010
16	0.172	0.036	+0.136	0.520 (−11.60)	0.641 (+0.50)	0.102	0.070
17	0.158	0.030	+0.128	0.558 (−7.80)	0.640 (+0.40)	0.080	0.078
19	0.094	0.028	+0.066	0.610 (−2.60)	0.639 (+0.30)	0.044	0.050
Mistral7B							
13	0.040	0.056	−0.016	0.674 (0.00)	0.680 (+0.60)	0.016	0.024
16	0.037	0.051	−0.014	0.681 (+0.70)	0.681 (+0.65)	0.017	0.020
17	0.048	0.045	+0.003	0.674 (−0.05)	0.711 (+3.65)	0.021	0.026
19	0.036	0.043	−0.007	0.672 (−0.20)	0.673 (−0.10)	0.015	0.021

Residual interventions produce more flips and lower accuracy than attention interventions at all four Qwen7B layers. Mistral7B shows a different component ordering. The direction counts separate harmful-to-benign from benign-to-harmful changes within each residual response.

Corrections and new errors. By Equation 5, the rate of corrected errors is $(\mathrm{FlipRate} + \Delta\mathrm{Accuracy})/2$ and the rate of new errors is $(\mathrm{FlipRate} - \Delta\mathrm{Accuracy})/2$. Table 10 computes both from the unrounded flip rates and accuracy changes of the settings in Table 9. At Qwen7B layer 16, residual noise corrects 2.80% of predictions but introduces errors on 14.40%, which gives its accuracy loss of 11.60 points. The attention comparison corrects 2.05% and introduces errors on 1.55%, a gain of 0.50 points. Accuracy depends on the balance between the two, so more flips need not give higher accuracy.

Table 10: Corrections and new errors in the component comparisons. Rates are percentages of all 500 inputs, averaged over the same four target settings as Table 9. Corrected: a wrong clean prediction that the noise makes correct. New error: a correct clean prediction that the noise makes wrong. The residual and attention columns use the same comparisons as that table.

	Residual		Attention	
Layer	Corrected	New error	Corrected	New error
Qwen7B				
13	0.60	12.60	0.75	2.35
16	2.80	14.40	2.05	1.55
17	4.00	11.80	1.70	1.30
19	3.40	6.00	1.55	1.25
Mistral7B				
13	2.00	2.00	3.10	2.50
16	2.20	1.50	2.85	2.20
17	2.35	2.40	4.05	0.40
19	1.70	1.90	2.10	2.20

D Perturbation Implementation Details

Table 11 specifies the intervention sites. Hidden state perturbations add Gaussian noise to block outputs, while the attention control rescales quantities returned by the self-attention module. Equal parameter values do not produce equal relative tensor changes at these sites, so component comparisons also use the observed changes in decision logits.

Table 11: Each operator changes a distinct internal site. Layer profiles use residual stream perturbations, with attention and MLP interventions as comparisons.

Component	Modified site	Token scope	Evaluation
Residual stream	Transformer block output	all prompt tokens	primary layer profile
Attention return	self-attention return	all prompt tokens	attention control
Attention branches	output, value, or logits	all prompt tokens	output-matched comparisons
MLP / residual input	module output or prehook	all prompt tokens	additional comparison

The implementation is hook based. Hidden state perturbations are registered on the selected layer output. Attention controls are registered on the self attention module and preferentially rescale returned attention weights; when the implementation does not expose weights directly, the hook rescales the returned attention tensor instead. Auxiliary analyses also support attention output, attention value, attention logits, MLP output, residual input, and value path perturbations. We use these variants as calibration checks when interpreting near zero attention flips.

E Calibration at Similar Output Changes

The Qwen0.5B component comparison uses one 500-example block, layers 13, 16, 17, and 19, and perturbation values 0.05/0.1/0.2. Because these values induce different relative tensor changes across components, we also group settings by the observed KL divergence and logit shift of the decision logits. This analysis includes 8 Qwen0.5B pairs and 64 Qwen7B/Mistral7B pairs; Table 9 compares the latter two models.

Table 13 compares component responses on Qwen2.5-0.5B-Instruct using a 500-example ToxicChat block, labels read from eight greedily decoded tokens, layers 13, 16, 17, and 19, and perturbation values 0.05, 0.10, and 0.20. Hidden-state flip rates are usually larger across these layers. Attention-logit perturbations also produce nonzero flip rates at layers 16 and 17, despite differences in relative tensor change across operators.

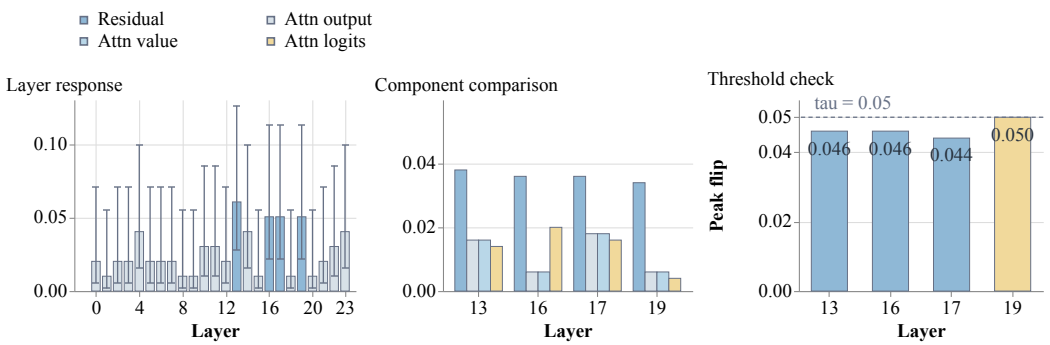

Figure 6: Component and depth structure in Qwen0.5B. Left: flip rates by layer with 95% Wilson intervals on 99 examples. Center: mean component responses over three magnitudes on a 500-example block. Right: peak flip rates in the magnitude sweep, with a reference flip rate of 0.05 shown as a dashed line.

Table 12: Hidden state interventions produce larger output movement than attention branches in the Qwen0.5B calibration pairs. Values are means after grouping runs by observed KL.

Path	Flip	KL	Logit shift
attention output	0.0076	0.0049	0.2365
attention value	0.0076	0.0049	0.2363
hidden-state	0.0253	0.1239	1.4843

Table 13: Flips under hidden state perturbations exceed flips under perturbations to attention branches at most tested layers and magnitudes. Results use one 500-example block.

Layer	Path	Flip 0.05	Flip 0.10	Flip 0.20	Rel. change 0.20
13	hidden-state	0.0280	0.0360	0.0380	0.2000
13	attention output	0.0160	0.0160	0.0140	0.8000
13	attention value	0.0160	0.0160	0.0140	0.8000
13	attention logits	0.0140	0.0140	0.0140	0.8000
16	hidden-state	0.0260	0.0360	0.0340	0.1999
16	attention output	0.0060	0.0060	0.0060	0.8000
16	attention value	0.0060	0.0060	0.0060	0.8000
16	attention logits	0.0200	0.0200	0.0140	0.8000
17	hidden-state	0.0240	0.0360	0.0320	0.1999
17	attention output	0.0180	0.0160	0.0160	0.8000
17	attention value	0.0180	0.0160	0.0160	0.8000
17	attention logits	0.0140	0.0140	0.0160	0.8000
19	hidden-state	0.0280	0.0340	0.0340	0.2000
19	attention output	0.0060	0.0060	0.0040	0.8000
19	attention value	0.0060	0.0060	0.0040	0.8000
19	attention logits	0.0040	0.0040	0.0040	0.8000

The dense magnitude sweep uses twelve magnitudes of Gaussian noise, $\delta \in \{0.001, 0.002, 0.005, 0.01, 0.02, 0.05, 0.1, 0.2, 0.5, 1, 2, 5\}$, on the same 500-example block. Table 14 includes all tested values for layers 13, 16, 17, and 19. Peaks stay at or below 0.0500, layer 17 remains below threshold with a peak of 0.044, and the curves are not monotone in the perturbation magnitude.

Table 15 compares positive and negative hidden state perturbations on the same four layers and 500 examples, with eight magnitudes per sign. Flip rates differ at some settings, including layer 17 at $\delta = 0.1$, but opposite signs yield similar responses across much of the tested range.

Table 14: Flip rates rise into a broad plateau rather than a sharp threshold. Entries report flip rate / accuracy on a 500-example block.

δ	Layer 13	Layer 16	Layer 17	Layer 19
0.001	0.0020 / 0.5000	0.0000 / 0.5020	0.0000 / 0.5020	0.0000 / 0.5020
0.002	0.0100 / 0.4960	0.0060 / 0.5000	0.0040 / 0.4980	0.0000 / 0.5020
0.005	0.0140 / 0.5040	0.0060 / 0.4960	0.0040 / 0.5020	0.0020 / 0.5000
0.010	0.0180 / 0.4920	0.0160 / 0.4940	0.0120 / 0.4940	0.0120 / 0.4940
0.020	0.0360 / 0.4940	0.0280 / 0.4900	0.0220 / 0.5000	0.0160 / 0.5020
0.050	0.0400 / 0.5100	0.0400 / 0.4980	0.0300 / 0.5040	0.0220 / 0.5000
0.100	0.0420 / 0.5040	0.0460 / 0.4920	0.0300 / 0.5120	0.0340 / 0.5040
0.200	0.0460 / 0.5200	0.0380 / 0.5120	0.0320 / 0.5100	0.0500 / 0.5040
0.500	0.0400 / 0.5140	0.0460 / 0.5160	0.0440 / 0.5260	0.0440 / 0.5100
1.000	0.0460 / 0.5080	0.0400 / 0.5020	0.0440 / 0.5260	0.0460 / 0.5200
2.000	0.0340 / 0.5200	0.0360 / 0.5180	0.0340 / 0.5200	0.0360 / 0.5140
5.000	0.0340 / 0.5200	0.0320 / 0.5180	0.0300 / 0.5200	0.0340 / 0.5160

Table 15: Changing perturbation sign leaves the broad layer pattern intact. Entries report negative / positive flip rates at matched magnitudes.

δ	Layer 13	Layer 16	Layer 17	Layer 19
0.001	0.0020 / 0.0020	0.0040 / 0.0000	0.0000 / 0.0000	0.0020 / 0.0000
0.005	0.0080 / 0.0060	0.0060 / 0.0120	0.0080 / 0.0020	0.0020 / 0.0080
0.010	0.0140 / 0.0240	0.0140 / 0.0080	0.0140 / 0.0140	0.0100 / 0.0100
0.050	0.0320 / 0.0360	0.0380 / 0.0260	0.0300 / 0.0260	0.0260 / 0.0180
0.100	0.0380 / 0.0340	0.0380 / 0.0320	0.0360 / 0.0440	0.0440 / 0.0320
0.200	0.0440 / 0.0500	0.0360 / 0.0500	0.0320 / 0.0340	0.0500 / 0.0420
0.500	0.0400 / 0.0400	0.0440 / 0.0500	0.0380 / 0.0360	0.0380 / 0.0320
1.000	0.0400 / 0.0420	0.0460 / 0.0420	0.0420 / 0.0420	0.0420 / 0.0380

Table 16 compares over-refusal and benign utility with the original prompts and with prompt template T0 (Figure 10) on Qwen0.5B and Qwen7B. These evaluations use three seeds. On OR-Bench, utility is the answer rate, $1 - \mathrm{ORR}$, where ORR is the over-refusal rate, so the two columns report the same outcome once as refusal and once as answering.

Table 16: T0 sharply raises over-refusal for Qwen0.5B but changes Qwen7B little. Rows average 60-example evaluations over three seeds.

Model	Dataset	ORR raw	ORR T0	Utility raw	Utility T0
Qwen0.5B	OR-Bench	0.0133	0.6425	0.9867	0.3575
Qwen0.5B	XSTest	0.0025	0.2458	0.5250	0.2817
Qwen7B	OR-Bench	0.0050	0.0225	0.9950	0.9775
Qwen7B	XSTest	0.0008	0.0050	0.5267	0.5225

F Additional Comparisons across Models

The matched Qwen comparison uses the same 99-example block, three depth-scaled layers (layers at the same relative depths in each model), and three perturbation magnitudes for every model from 1.5B to 72B. Table 18 also shows responses over grids of different sizes. The combination counts specify the extent of each sweep.

Mean flip rates in Table 18 depend on which layers and magnitudes are included. Comparisons across model sizes therefore use the common grid in Table 17, rather than ranking models by averages over different grids.

Table 17: Sensitivity across Qwen sizes on the same nine depth-scaled combinations of layer and noise settings and 99 examples. Bootstrap intervals describe mean flip rate across the nine settings.

Model	Mean flip	95% interval	Max flip	Accuracy change
Qwen1.5B	0.0011	[0.0000, 0.0034]	0.0101	−0.0011
Qwen3B	**0.0281**	[0.0168, 0.0393]	**0.0606**	−0.0101
Qwen7B	0.0090	[0.0034, 0.0146]	0.0202	−0.0045
Qwen14B	0.0224	[0.0180, 0.0269]	0.0303	**+0.0180**
Qwen32B	0.0180	[0.0112, 0.0236]	0.0303	−0.0045
Qwen72B	0.0067	[0.0022, 0.0112]	0.0202	+0.0045

Table 18: Broader Qwen sweeps show no monotonic relation between scale and local sensitivity. These runs cover different layer and magnitude combinations.

Model	Combinations	Mean flip	Flip range	Mean pert. acc.
Qwen0.5B	12	0.0236	0.0101–0.0404	0.5572
Qwen1.5B	9	0.0022	0.0000–0.0101	0.5657
Qwen3B	9	0.0505	0.0101–0.0707	0.5511
Qwen7B	9	0.0135	0.0000–0.0303	0.5578
Qwen14B	32	0.0518	0.0000–0.0808	0.5600
Qwen14B	55	0.1000	0.0000–0.2500	0.4705
Qwen32B	9	0.0247	0.0000–0.0505	0.5623
Qwen32B	35	0.0143	0.0000–0.1250	0.4857
Qwen72B	4	0.0000	0.0000–0.0000	1.0000

G Decision Extraction and Judge Control

The model generates a short completion with greedy decoding, which is mapped to a binary label by the lexical rule in Table 20. Applying the same rule before and after perturbation ensures that label changes come from the model's outputs and not from the extraction rule; Table 8 tests whether three alternative judge phrasings preserve the resulting labels.

H Statistical Estimation

We use binomial intervals for flip rates on individual data blocks, bootstrap intervals for aggregates across layers and magnitudes, and variation across seeds for evaluations of prompt formatting. Each estimate reflects the unit over which the corresponding result is averaged.

Pointwise flip rates. Each 99-example block yields a binomial flip count, so we interpret pointwise uncertainty with Wilson intervals. For example, the 0.0505 flip rate points in the directional sweep correspond to Wilson 95% intervals of approximately [0.0218, 0.1128], while the 0.0404 points in the magnitude sweep correspond to approximately [0.0158, 0.0993]. The intervals leave the ordering in the threshold plots unchanged, and they confirm that the observed movement is small in absolute terms.

Cross model summaries. Cross model means are computed over the measured layer and magnitude combinations for each model, and uncertainty is summarized with nonparametric bootstrap intervals over those combinations. Each model uses the same three depth-scaled layers and three perturbation magnitudes. The intervals in Table 17 describe variation across these nine settings and not uncertainty over independently sampled models.

Variation across seeds. For evaluations of prompt formatting with three seeds, we report the mean and standard deviation across seeds. These summarize variation under the same model and dataset. Table 16 compares the original prompts with T0, and Figure 10 compares four templates.

I HUMAN REVIEW OF RESPONSE CHANGES

Human review separates substantive decision changes from changes in wording. We sample responses from attention controls, near-threshold settings, highly responsive layers, and larger perturbations, including layer 17. Each response pair contains the user input, baseline and perturbed outputs, extracted labels, and correctness relative to the dataset label.

Table 19: Readout changes across targeted perturbations. H/B denote harmful/benign predictions, not ground-truth labels; accuracy change is relative to baseline. The same 500 inputs are reused in each setting.

Group	Settings	Flips	H to B	B to H	Flip rate range	Mean accuracy	Accuracy change
Attention	2	11	3	8	0.006–0.016	0.4990	−0.0030
Residual (16)	3	52	28	24	0.010–0.056	0.5087	+0.0067
Residual (13/17/19)	5	94	61	33	0.030–0.048	0.5140	+0.0120

The attention controls, residual interventions at layer 16, and residual interventions at layers 13, 17, and 19 contain 1,000, 1,500, and 2,500 pairs of inputs and settings, respectively. Their perturbed accuracy ranges are 0.496–0.502, 0.496–0.526, and 0.508–0.524; ranges of accuracy change are -0.006–0.000, -0.006–0.024, and +0.006–0.022.

The residual interventions at layers 13, 17, and 19 produce 94 flips in 2,500 pairs of inputs and settings, compared with 11 in 1,000 pairs from attention controls. Their mean accuracy change is also positive rather than negative. The selected layers thus change the readout more often, and the direction counts show that flip rate alone does not measure quality.

I.1 SAMPLING AND ANNOTATION

We generate baseline and perturbed responses from Qwen2.5-0.5B-Instruct on the same 500 ToxicChat inputs under ten perturbation settings. For each input–setting pair, we retain both responses, their binary predictions, flip direction, and correctness. Sampling up to 100 flipped and 100 non-flipped records with random seed 42 yields 200 response pairs for review. Manual reviewers applied the rubric with three categories below; the agreement values summarize their labels.

Table 20: Semantic review preserves both generations and assigns one constrained label to each pair.

Field	Configuration
Model and data	Qwen2.5-0.5B-Instruct on a 500-example ToxicChat block; each setting uses the same inputs.
Baseline and readout	Baseline generations are recomputed before perturbation; the binary harmful label is 1 when the generated text contains “yes” or “harmful”, and 0 otherwise.
Perturbation settings	Ten settings: two attention controls and eight residual stream interventions. Their layers and magnitudes appear in Table 22.
Recorded fields	Setting, perturbation type, layer, value, sample index, conversation id, user input, ground truth, baseline and perturbed text, binary labels, flip flag, flip direction, and correctness fields.
Sampling	Up to 100 flipped and 100 non-flipped records sampled with random seed 42; the reviewed set contains 200 records.
Manual review	Reviewers assigned one label from the rubric with three categories below to each response pair; agreement measures consistency between their labels.

Reviewers assign exactly one of three interpretation categories:

- **meaningful decision change**. The perturbed output crosses the harmful or benign decision boundary under the rubric in a way that changes the task level judgment rather than only the surface form;
- **stylistic variation**. The perturbed output changes wording, fluency, or truncation without changing the substantive judgment;
- **judge ambiguous**. The output is too terse, malformed, or sensitive to the judgment prompt to support a confident semantic interpretation.

Agreement is computed over the three category labels, without scoring free-form explanations.

Table 21: Manual agreement reaches 0.955 overall and is lowest for attention controls, where sparse changes are most ambiguous.

Response group	Records	Agreement	Interpretation
Attention controls	26	0.846	Lowest agreement; small output changes are more difficult to classify.
Residual (16)	78	0.949	Reviewers generally agree on changes near the response threshold.
Residual (13/17/19)	96	0.990	High agreement on responses from highly sensitive layers and larger perturbations.
All reviewed pairs	200	0.955	Overall agreement between manual reviewers.

Table 22 reports all evaluated pairs of inputs and settings, whereas manual agreement above describes only the reviewed subset.

Table 22: Flip directions for all ten evaluated perturbation settings. Rows provide the results underlying Table 19.

Setting	Layer	Magnitude	n	Flips	Flip rate	Harmful to benign / Benign to harmful
Attention return	19	1.50	500	8	0.0160	3 / 5
Attention return	16	1.50	500	3	0.0060	0 / 3
Residual stream	16	0.01	500	5	0.0100	2 / 3
Residual stream	13	0.20	500	18	0.0360	13 / 5
Residual stream	13	0.50	500	24	0.0480	15 / 9
Residual stream	17	0.20	500	18	0.0360	12 / 6
Residual stream	17	0.50	500	19	0.0380	13 / 6
Residual stream	19	0.10	500	15	0.0300	8 / 7
Residual stream	16	0.10	500	28	0.0560	14 / 14
Residual stream	16	0.20	500	19	0.0380	12 / 7

J Additional Limitations

The matched Qwen comparison uses a common data block, three depth-scaled layers, and three perturbation magnitudes, but covers only one model family. Residual and attention interventions are compared through their downstream effects, not through equal changes at the intervention site. Prompt wording and label extraction can also affect measured sensitivity. Human review covers a sampled set of response pairs rather than every evaluated output.

Output validity limits the generated-label cross-model comparison. TinyLlama produces labels compatible with the original threshold evaluation, whereas the high invalid-output rates for GPT2 and Llama prevent a comparable analysis under that prompt and extraction rule. The official-chat analyses in Appendices A and B instead retain continuous class margins and report unrestricted top-token validity separately. Combined hidden-state and attention perturbations also differ in strength from either perturbation alone, so their flip rates do not isolate a benefit from combining components. Responses vary by task: toxicity and bias evaluations show changes, while the evaluated jailbreak examples do not.

K Layer Responses and Selection on Later Inputs

Response patterns across input blocks. With the same generated-judgment labels and perturbation magnitude, all five input blocks show a responsive region in early layers, although the exact maximum moves. Figure 7 shows the full profiles. We therefore test whether responsive regions can guide measurements on later inputs.

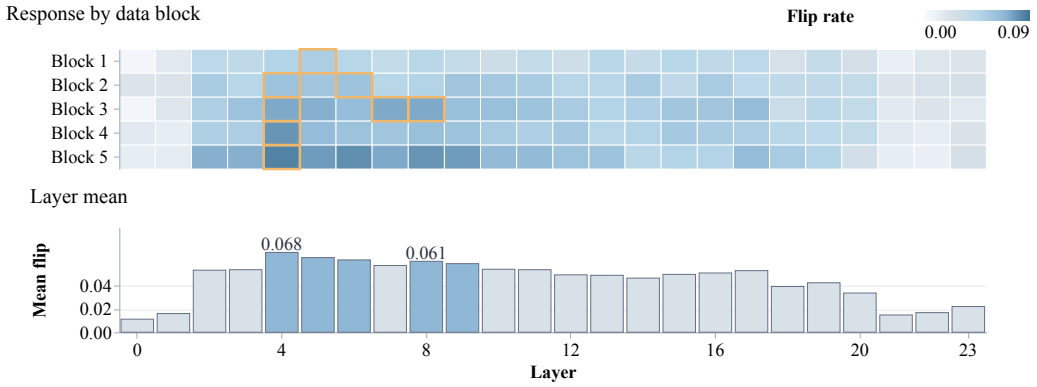

Figure 7: Response patterns repeat across input blocks. Cells show flip rates by layer for five consecutive, non-overlapping 500-example ToxicChat blocks on Qwen0.5B at magnitude 0.1; bars summarize their mean. A shared early region remains responsive despite changes in the maximizing layer.

Selecting on four blocks and evaluating on the fifth separates localization from evaluation. A two-layer window contains a response maximum in all five folds. One selected layer retains 98.5% of the exhaustive peak, compared with 65.4% for a random layer. Top-ranked individual layers tie the contiguous window at these budgets, so the benefit comes from reusing response information and not from adjacency as such.

Selecting layers under a budget. A layer profile can also guide a subsequent search. Let $s_{\ell,k}$ be the measured response at layer ℓ in profiling block k. Let $\mathcal{T}$ index the profiling blocks. We average their responses, $\bar{s}_\ell = |\mathcal{T}|^{-1} \sum_{k \in \mathcal{T}} s_{\ell,k}$, without using the held-out block. For a budget of b layers, let $\mathcal{W}_b$ contain all contiguous windows of length b. We select

$$W_b^\star = \arg\max_{W \in \mathcal{W}_b} \frac{1}{b} \sum_{\ell \in W} \bar{s}_\ell. \tag{10}$$

Ties are resolved by the earliest starting layer. A contiguous window keeps a local region and does not assume that the exact peak stays in place across examples. The selected layers define the candidate set for evaluating the next input block.

Evaluating the selected window. On held-out block k, let $m_k = \max_\ell s_{\ell,k}$ be the exhaustive peak and $q_k(W) = \max_{\ell \in W} s_{\ell,k}$ the best response in the selected window. We measure

$$\text{Coverage}_k(W) = \mathbf{1}[q_k(W) = m_k], \qquad \text{Retention}_k(W) = \frac{q_k(W)}{m_k}. \tag{11}$$

Retention is defined as one when every layer has zero response. Coverage accepts any tied maximum; retention measures how much response remains when the exact maximum is missed. We average both quantities across held-out blocks. Window selection uses only the other blocks, including all tie-breaking decisions. The held-out maximum therefore evaluates a choice made without seeing that block. The budget counts candidate layers on the held-out block; the initial profile has a separate cost.

Cost of reusing layer measurements. Let K profiling blocks and M future blocks each contain N inputs, with L candidate layers and selected budget b. Counting a shared baseline and one

Table 23: Selection based on measured responses outperforms rules based only on position at the same budget. Contiguous windows selected by mean response and top-ranked individual layers tie at both budgets. Coverage and retention are percentages averaged over five held-out blocks.

Selection rule	One layer		Two layers	
	Coverage	Retention	Coverage	Retention
Window with highest mean	**80.0**	**98.5**	**100.0**	**100.0**
Top-ranked layers	**80.0**	**98.5**	**100.0**	**100.0**
First layers	0.0	16.3	0.0	25.3
Middle window	0.0	77.8	0.0	79.4
Even spacing	0.0	70.5	20.0	89.9
Random window	6.7	65.4	13.0	73.6

perturbation per candidate gives

$$C_{\text{full}} = MN(L+1), \qquad C_{\text{profile}} = KN(L+1) + MN(b+1). \tag{12}$$

The initial scan pays off when $M > K(L+1)/(L-b)$. Four profiling blocks, 24 layers, and a two-layer budget break even after five equally sized future blocks. Once a profile exists, a new block requires three evaluations per input rather than 25. The calculation counts model evaluations; runtime also depends on the cost of each evaluation.

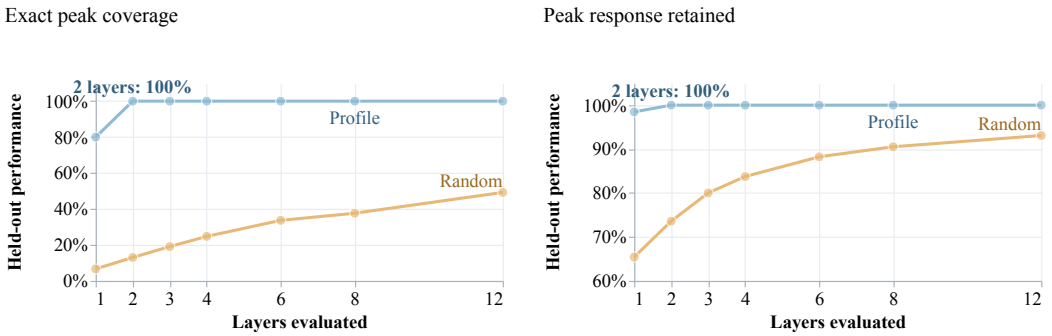

Figure 8: Two profiled layers recover a held-out maximum in all five blocks. Each window is selected on four blocks and evaluated on the fifth; the random baseline averages all contiguous windows of the same size. One profiled layer retains 98.5% of peak sensitivity, compared with 65.4% for a random layer. Complete values appear in Table 24.

The layer-budget experiment uses the five non-overlapping ToxicChat blocks in Figure 7, with hidden state perturbation magnitude $\delta = 0.1$ at every layer. For each budget, the contiguous window with the highest mean flip rate on four blocks is evaluated on the remaining block. The random baseline averages all possible contiguous windows with that budget. Coverage is the fraction of folds in which the selected set contains a held-out maximum; retention is the selected set's best held-out flip rate divided by that maximum.

L Prompt Formatting and Behavioral Tradeoffs

The matrix below compares attack success with behavior on benign requests under the same prompt templates. On OR-Bench, the utility panel shows the answer rate, $1 - \text{ORR}$, so a drop in utility equals the rise in over-refusal. Reading the harmful and benign results together shows whether lower attack success comes with a loss of ordinary answering.

M Additional Models and Output Validity

TinyLlama outputs can be parsed using the same label rule as the Qwen threshold evaluation. The comparison uses the same 99 examples, three depth-scaled layers on the 22-layer backbone (6, 11,

Table 24: Two profiled layers recover every held-out maximum; random selection reaches only 13.0% coverage at the same budget. Values are percentages.

Layers	Window with highest mean		Random selection	
	Coverage	Retention	Coverage	Retention
1	80.0	98.5	6.7	65.4
2	100.0	100.0	13.0	73.6
3	100.0	100.0	19.1	80.0
4	100.0	100.0	24.8	83.8
6	100.0	100.0	33.7	88.3
8	100.0	100.0	37.7	90.5
12	100.0	100.0	49.2	93.1

A lower attack success rate helps only if ordinary requests are still answered. Figure 9 places safety gains and utility losses on the same scale in percentage points, pairing the two model sizes within each benchmark. Qwen2.5-0.5B gains 53.8 points on AdvBench and 23.4 on JailbreakBench, but loses 58.6 points of utility on OR-Bench and 21.2 on XSTest.
Every reported 7B change is within 1.5 points. A taller bar is desirable for the harmful-prompt comparisons and undesirable for benign utility.

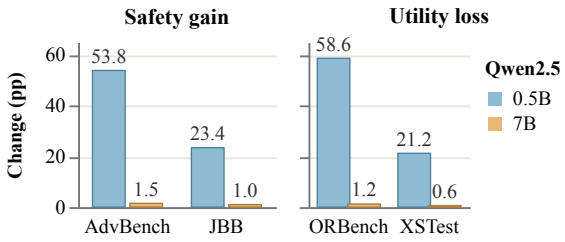

Figure 9: Safety gains and utility losses from prompt formatting (percentage points). Qwen2.5-0.5B changes strongly on both harmful prompts (left) and benign prompts (right); every 7B change stays within 1.5 points. JBB denotes JailbreakBench.

Read together, the panels show that the small model's lower attack success comes with a large loss of ordinary answering, so it is not a selective improvement in safety.

16), and the noise grid 0.05/0.1/0.2. The positive and negative direction sweeps reach peak flip rates of 0.0202 and 0.0101, respectively. The magnitude sweep peaks at 0.0303 and does not exceed a flip rate of 0.05.

Table 25 reports generated-label validity under the original evaluation prompt. Invalid-output rates are high for GPT2 and Llama, preventing comparison with TinyLlama under that extraction rule. These results are separate from the official-chat class-logit measurements used for the study of output changes.

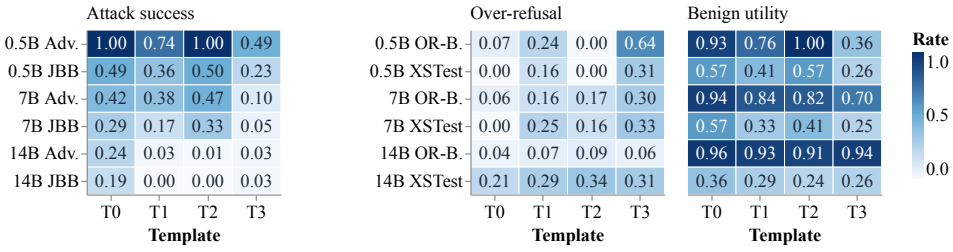

Figure 10: Prompt templates move harmful and benign behavior together. Columns are templates T0–T3. Left: attack success on AdvBench and JailbreakBench for three Qwen sizes (lower is better). Center: over-refusal on OR-Bench and XSTest for the same models (lower is better). Right: benign utility on the same combinations of models and datasets (higher is better). No template dominates all three outcomes.

Table 25: TinyLlama outputs support the threshold comparison. High invalid-output rates prevent the same comparison for GPT2 and Llama.

Model	Response and output validity	Comparison
TinyLlama	Direction peak 0.0202; peak flip rate 0.0303; stable extraction	compatible
GPT2-XL	Invalid outputs: 87.00% toxic / 85.47% benign	excluded
Llama-2-7B	Invalid outputs: 43.30–91.29% toxic / 44.34–92.55% benign	excluded
Llama-3.1-8B	Invalid outputs: 25.20–65.55% toxic / 20.13–60.28% benign	excluded

N Models, Data, and Reproducibility

The experimental descriptions and result tables cover perturbation sweeps, model comparisons, review of response pairs, evaluations of prompt templates, comparisons of judge phrasing, and interpretation of continuous scores.

Models. Qwen2.5-Instruct models are obtained from `https://huggingface.co/collections/Qwen/qwen25`; the 0.5B, 1.5B, 7B, 14B, and 32B releases use Apache-2.0, while 3B and 72B follow their terms in their model cards. Mistral-7B-Instruct-v0.2 and TinyLlama-1.1B-Chat-v1.0 are released under Apache-2.0. Llama-3.1-8B-Instruct follows the Llama 3.1 Community License at `https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct`.

Datasets. AdvBench uses the MIT-licensed `llm-attacks` repository. JBB-Behaviors is MIT licensed; OR-Bench and XSTest are CC-BY-4.0; ToxicChat is CC-BY-NC-4.0. Derived blocks follow the corresponding source terms.

O Further Context and Related Work

Safety training and adversarial evaluation. Constitutional AI trains assistants with feedback based on a set of written principles (Bai et al., 2022). The Jailbroken study analyzes how competing training objectives and gaps in generalization can undermine safety training (Wei et al., 2023). Red teaming by humans (Ganguli et al., 2022), test cases written by another language model (Perez et al., 2022), and iterative attacks such as PAIR (Chao et al., 2023) find failures in the generated responses. JailTrickBench studies how attack settings affect comparisons between attacks (Xu et al., 2024), and rewriting a request in another tense can expose failures from such a small change (Andriushchenko & Flammarion, 2024). These studies show that it is important to test what an intervention changes. Our hidden-state experiments keep the input text fixed and trace how the decision responds to the noise. The prompt-formatting experiments are a separate comparison of generated behavior.

Steering and the structure of refusal. Activation addition and representation engineering control outputs by adding directions to hidden states (Turner et al., 2023; Zou et al., 2023a), and contrastive activation addition applies this approach to several behaviors (Panickssery et al., 2024). Studies of refusal have found a single direction with a strong effect on refusal (Arditi et al., 2024), several independent directions and concept cones (Wollschläger et al., 2025), and representations of harmfulness that are separate from those of refusal (Zhao et al., 2025). Steering Externalities examines how directions derived from benign tasks can increase jailbreak risk (Xiong et al., 2026), and Malla et al. show that steering along random and learned directions can cause similar behavioral tendencies (Malla et al., 2026). We relate the noise allocation to how much the response depends on the decision gradient. We also relate the token positions that pass the response on to the output changes that predict individual judgments.

Gradients and safety boundaries. Gradient Cuff uses the refusal loss and its gradient norm to detect jailbreak attempts (Hu et al., 2024). X-Boundary separates harmful representations from nearby safe ones, which improves defense while keeping the model useful (Lu et al., 2025). RASS samples representations near a safety boundary to study and reduce over-refusal (Pan et al., 2025). These works relate the structure of representations near a boundary to safety outcomes. Our comparisons

that remove the gradient component ask which parts of the noise produce a measured response when the same total energy is divided among tokens in different ways. The labeled comparisons then show how many of the resulting flips correct errors and how many introduce new ones.

Interventions and causal localization. Causal abstraction tests whether interventions on neural representations agree with interventions in an interpretable causal model (Geiger et al., 2021). Causal tracing locates the states involved in recalling facts (Meng et al., 2022), and path patching tests hypotheses about which paths through the model carry a behavior (Goldowsky-Dill et al., 2023). The choice of corruption method and evaluation metric can substantially change the conclusions of activation-patching experiments (Zhang & Nanda, 2024). Our position comparisons keep the original noise values at the selected positions and do not rescale them. Each kept part therefore applies the same noise as in the full perturbation, while we test how earlier and final prompt states affect the decision score. This complements analyses of the pathways that carry refusal steering (Cheng et al., 2026).

Decoding representations and predicting responses. Control tasks help separate the information contained in a representation from what the probe itself learns (Hewitt & Liang, 2019). Consistency constraints without labels can recover a model's latent answers from its activations (Burns et al., 2022), and the tuned lens uses affine probes to decode intermediate states into distributions over the vocabulary (Belrose et al., 2023). Our prediction test instead fits the change in the final output logits of tokens other than the class answers, and then predicts class margins on other inputs and noise directions. This links a change across the whole vocabulary to individual safety judgments, and the class margins serve as a separate prediction target.

Prompt variation, judges, and uncertainty. Part of the apparent sensitivity to prompts can come from rigid answer matching and other evaluation choices (Hua et al., 2025). Controlled paraphrases based on how users actually phrase requests offer another way to audit the effect of wording (Chataigner et al., 2026). Hypothesis tests on output distributions separate changes caused by an intervention from ordinary variation in the outputs (Rauba et al., 2025). Studies of LLM judges also find position, verbosity, and self-enhancement biases (Zheng et al., 2023). For these reasons, we state the readout, the inputs, and the noise comparison explicitly. Our paired noise signs, alternative forms of the label, held-out predictions, and semantic checks each address a different part of this measurement problem.

Data and model sources. ToxicChat provides toxicity labels for real user queries to a chatbot, which include difficult cases that toxicity data from social media do not cover (Lin et al., 2023). OR-Bench and XSTest add benign requests that may cause excessive refusal (Cui et al., 2024; Röttger et al., 2024). The model families used in our comparisons are documented in the Qwen2.5, Llama 3, and Mistral reports (Qwen et al., 2024; Grattafiori et al., 2024; Jiang et al., 2023). These sources describe the data and models. All response sizes, position effects, and prediction results reported here come from our own experiments.