RETHINKING REFUSAL DIRECTIONS: IDENTIFICATION AND INTERVENTION EFFECTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Low-dimensional activation directions can causally control refusal in large language models, but a direction's identity and behavioral meaning are often treated as properties of the vector alone. We show that neither is fixed. First, the direction identified as controlling refusal depends on how candidates are constructed and evaluated: lexical and activation-space objectives can select closely aligned or substantially different directions. A causally effective direction is therefore better understood as an operational causal approximation than as a uniquely recovered internal axis. Second, full-generation semantic judgments analyzed with Signal Detection Theory show that harmful and benign refusal rates change differently across operators, models, and prompt populations, producing changes in both response criterion c and the descriptive probit-scale separation index d'. Third, in a separate continuous-score analysis of contrastive activation addition, we identify a locally stable prompt-specific downstream gain G_i. Item-specific gains reconstruct local proxy-discrimination trajectories more accurately than class-level or global response models, demonstrating that a common activation displacement can produce heterogeneous downstream score effects across prompts. Together, these results show that a refusal direction functions as an operational causal interface whose behavioral meaning emerges jointly from its identification procedure, intervention operator, prompt state, downstream computation, and evaluation population.

1 INTRODUCTION

Mechanistic interpretability seeks to understand how internal representations and computations in large language models give rise to observable behavior. A growing line of work studies these mechanisms through causal interventions on model activations, identifying and manipulating low-dimensional directions that control specific behaviors Turner et al. (2023); Rimsky et al. (2024). Prior work has shown that refusal can be causally mediated by a single activation direction across multiple models and safety-aligned languages Arditi et al. (2024); Wang et al. (2025). Other work challenges the interpretation of this direction as a unique internal axis, instead describing refusal as a multidimensional concept cone containing multiple mechanistically independent directions Wollschläger et al. (2025). Work on exaggerated safety behavior further shows why overall refusal rate is insufficient as an evaluation metric: it does not distinguish selective refusal of harmful inputs from indiscriminate refusal of benign inputs Röttger et al. (2024).

We therefore take an operational view of refusal directions. Rather than asking whether a particular vector is "the" internal representation of refusal, we ask when a constructed direction provides an adequate causal approximation of refusal behavior, what determines which direction is identified, and what an intervention along that direction actually changes. To answer these questions, we audit key choices in direction construction and decompose intervention-induced behavior into distinct components.

We model refusal as a two-class decision process using Signal Detection Theory (SDT) Green & Swets (1966); Macmillan & Creelman (2005). Harmful prompts are signal trials and benign prompts are noise trials; semantic refusal is the positive response and compliance is the negative response. A refusal to a harmful prompt is therefore a hit, compliance with a harmful prompt is a miss, refusal to a benign prompt is a false alarm, and compliance with a benign prompt is a correct rejection.

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.

This formulation separates two aspects of refusal behavior that overall refusal rate conflates. Rate-based d' is a probit-scale index of harmful–benign refusal separation, whereas c summarizes general response tendency. Their latent sensitivity and criterion interpretation requires equal-variance evidence; with one operating point per condition, we use them descriptively.

To examine how direction construction and intervention design affect these behavioral dimensions, we study two model families at two scales: Gemma 2–2B-IT, Gemma 2–9B-IT, Qwen 2.5–3B-Instruct, and Qwen 2.5–7B-Instruct Gemma Team et al. (2024); Qwen Team et al. (2024). We compare multiple refusal-direction construction procedures and evaluate three intervention operators. Each model is tested on balanced harmful and benign subsets of XSTest Röttger et al. (2024) and WildGuardTest Han et al. (2024) using full-generation semantic-refusal judgments. Interventions are restricted to prompt prefill, allowing us to measure how modifying the prompt representation changes subsequent refusal behavior without repeated generation-time forcing.

Our analysis revises the usual interpretation of refusal directions at three levels. First, direction identity is operational: the vector recovered as r depends on the construction and selection procedure. Second, the functional effect of acting on r is state- and operator-dependent: the same geometric displacement can produce different downstream responses across prompts. Third, the resulting behavior is multidimensional: interventions alter harmful and benign refusal differently rather than controlling only an overall refusal rate. We establish the behavioral result using full-generation semantic judgments and analyze one local downstream mechanism in a separate continuous first-token proxy system.

2 Related Work

Refusal direction identification. Most refusal-direction methods construct candidate vectors from contrastive activations of harmful and benign prompts, but differ in how candidates are selected. Arditi-style methods use output-level refusal proxies Arditi et al. (2024), while Wang et al. allow model-specific refusal-token sets identified through pilot generations Wang et al. (2025). COSMIC instead removes dependence on predefined refusal tokens by using an activation-space selection objective Siu et al. (2025). Meanwhile, work on concept cones suggests that refusal need not correspond to a unique internal axis Wollschläger et al. (2025). We therefore treat direction identity as operational rather than canonical: we audit how alternative proxies and selection objectives change the selected r, and separately test its causal effect on full-generation semantic refusal.

Activation steering and intervention geometry. Contrastive activation addition (CAA) manipulates model behavior by adding a constant displacement along a steering direction Turner et al. (2023); Rimsky et al. (2024), whereas Arditi et al. use repeated directional ablation across layers to test the causal role of a refusal direction Arditi et al. (2024). More general affine and unified steering frameworks show that addition, scaling, and projection-based interventions can be described within a common geometric formalism; this operator unification is therefore not our contribution Marshall et al. (2024); Im & Li (2025). Recent methods such as CAST and AlphaSteer instead introduce conditional or input-dependent interventions to mitigate benign over-refusal Lee et al. (2025); Sheng et al. (2026). AlphaSteer learns layer-specific transformations rather than manipulating one fixed global direction. Work on self-repair shows that ablating an internal component can trigger compensatory downstream changes, first characterized as the Hydra effect and later documented across model families with substantial prompt-level variation McGrath et al. (2023); Rushing & Nanda (2024). Single-site directional scaling places our intervention in the same general perturbation–response setting, but we do not separately estimate the direct perturbation effect and subsequent compensation required to identify self-repair. In this work, we use CAA, a controlled single-site directional-scaling operator, and all-layer directional ablation to ask a different question: how does manipulation of a fixed refusal direction affect observed harmful and benign refusal rates, their probit-scale separation index, and criterion? CAA and single-site scaling admit a common local form within $\mathrm{span}(r)$, while all-layer ablation is treated separately as a repeated trajectory-level intervention.

Signal detection and local response models. Signal Detection Theory originated as an application of statistical decision theory to detection and discrimination, separating the quality of available evidence from the decision rule used to convert that evidence into a response Green & Swets (1966);

Macmillan & Creelman (2005). This distinction is useful for activation control because changes in the overall propensity to emit a binary behavior need not coincide with changes in how selectively that behavior is expressed across stimulus classes. Under classical equal-variance Gaussian SDT, these effects are summarized by response criterion c and sensitivity d', respectively; invariance of the traditional d' to response bias requires this restricted evidence model Aujla (2023). Our semantic judgments provide only one binary operating point per condition, so we use c and d' as descriptive behavioral indices rather than fully identified latent parameters. This decomposition nevertheless constrains intervention effects more strongly than overall refusal rate alone. Recent work has applied SDT to factual discrimination, showing that model temperature can affect both criterion and sensitivity Cacioli (2026), and to models' introspective predictions of their own refusal behavior Gondil (2026). Separately, activation-steering work has characterized local linear and Jacobian-based responses to activation perturbations Skifstad et al. (2026). To our knowledge, we are the first to use SDT to decompose full-generation semantic refusal under causal activation interventions and to pair this behavioral analysis with prompt-specific downstream responses measured in a distinct continuous proxy-score system.

3 Operationalizing Refusal Directions

A refusal direction is not specified by an activation contrast alone. Its identity also depends on how candidate directions are constructed, what evidence is used to select among them, and whether the selected direction causally controls the semantic behavior of interest. We therefore distinguish three stages—candidate construction, proxy-based selection, and semantic qualification—and use a same-bank COSMIC comparison as a separate cross-objective audit of the selection stage. This separation allows us to locate where agreement or disagreement enters the identification pipeline rather than treating the resulting vector as a uniquely defined representation. Exact proxy vocabularies, filtering rules, and selection procedures are provided in Appendix A.

3.1 Lexical-proxy sensitivity

We first test how strongly direction identification depends on the lexical evidence used to define refusal. Holding construction prompts, validation data, and cached intervention outputs fixed within each comparison, we vary the refusal-token proxy used during candidate filtering or ranking. The audit therefore isolates one source of operational dependence: the same underlying activation evidence can produce different admissible candidate banks or different rankings when "refusal evidence" is defined differently.

For Gemma 2, we cross candidate banks filtered by `I` or the combined `I+It` mass with selection scores based on the same two proxies. For Qwen 2.5, we hold an Arditi-prompt candidate bank filtered by the Wang-style `I+As` mass fixed and compare three lexical scores: `I`, `I+As`, and the literal control `I+It`. These diagnostic winners are used only for this exploratory pre-experiment setup; we subsequently adopt Wang's method for the formal intervention experiments.

The resulting checkpoint-dependent sensitivity is reported with the other direction-identity results in Section 5.1.

3.2 Formal direction construction

For the formal experiments, we construct one refusal direction independently for each checkpoint using the Wang-style English harmful–harmless prompt contrast Wang et al. (2025). Candidate directions are harmful-minus-harmless mean activation differences computed across candidate layer–position sites. The Wang procedure ranks this model-specific bank using a family-specific first-token refusal score and associated intervention checks, then selects one direction per checkpoint.

The direction is selected according to the construction prompts, refusal-token sets, and Wang selection procedure described in Appendix A, then fixed for all subsequent analyses.

3.3 Cross-objective audit with COSMIC

A direction preferred under one refusal objective need not be preferred under another. To test this without changing the candidate space, we apply COSMIC's activation-space objective Siu et al. (2025) to the same formal candidate bank and fixed prompt manifests used by the Wang procedure. We compare the objectives through winner identity, the frozen Wang direction's COSMIC rank, cosine similarity between their winners, and full-bank rank agreement.

COSMIC is used as a cross-objective audit rather than as an alternative ground truth. Disagreement indicates that the direction selected from a fixed bank depends on how refusal is operationalized; it does not establish that either objective uniquely recovers the underlying representation. The resulting four-model comparison is reported in Section 5.1.

3.4 Semantic qualification and freezing

Before formal SDT evaluation, we qualified each frozen direction using full generations under baseline and signed interventions. Qualification asked whether movement along the selected direction changed substantive refusal behavior rather than merely increasing a preferred first-token probability, producing a refusal-like opening followed by compliance, or degrading generation quality. These qualification runs used persistent generation-time intervention and were separate from the formal prefill-only experiment.

Semantic qualification served only as a validity check: directions were not selected or reselected from these outcomes, and the qualification runs do not enter the reported statistical analyses. Formal intervention effects are evaluated in Section 5.

4 Intervention and Measurement Framework

4.1 Experimental design and identification strategy

Our intervention design separates the geometry of a control direction from the behavioral effect of the operator applied to it. For each checkpoint, we freeze one refusal direction before the main experiments and vary the intervention operator while holding that direction fixed. This allows us to test whether acting on the same nominal low-dimensional control object produces a common behavioral effect, or whether the effect depends on the operator, model, and prompt population.

We evaluate four instruction-tuned checkpoints from two model families and two parameter scales: Gemma 2–2B-IT, Gemma 2–9B-IT, Qwen 2.5–3B-Instruct, and Qwen 2.5–7B-Instruct. Each model is tested on balanced subsets of XSTest Röttger et al. (2024) and WildGuardTest Han et al. (2024), with 150 harmful and 150 benign prompts per dataset, model, and condition. We analyze the datasets separately because they define distinct prompt populations and induce substantially different baseline refusal regimes. This separation tests whether an intervention effect depends on the population's operating regime rather than only on the model or direction.

We compare three ways of acting on the frozen direction: constant activation addition, single-site directional scaling, and repeated all-layer directional ablation. The first two are single-site writes within $\mathrm{span}(r)$; the third repeatedly modifies the residual-stream trajectory across layers. All interventions operate at valid prompt positions during prefill and are disabled during autoregressive decoding. The resulting behavior therefore reflects the downstream consequences of modifying the prompt representation, without persistent intervention on subsequently generated tokens.

The ten experimental conditions comprise baseline; CAA at $\alpha \in \{-1, -0.5, 0.5, 1\}$; single-site directional scaling at $\lambda \in \{0.25, 0.5, 0.75, 1\}$; and all-layer directional ablation. Generations use greedy decoding with at most 512 new tokens.

4.2 Intervention geometry

Let r be the unit-normalized frozen direction and h_{it} the residual-stream state at the selected layer for prompt i and valid prompt position t. Both single-site operators write pointwise along r:

$$h_{it}'^{(m)} = h_{it} + j_{it}^{(m)}, \qquad j_{it}^{(m)} = k_{it}^{(m)} r. \tag{1}$$

For CAA, the implemented coefficient α multiplies the raw Wang direction r_{raw}. Writing the intervention in the unit direction $r = r_{\mathrm{raw}}/\|r_{\mathrm{raw}}\|_2$ gives a constant geometric displacement k across prompts and positions:

$$k_{it}^{\mathrm{CAA}} = k = \alpha\|r_{\mathrm{raw}}\|_2, \qquad j_{it}^{\mathrm{CAA}} = kr = \alpha r_{\mathrm{raw}}. \tag{2}$$

For directional scaling, the coefficient depends on the existing refusal-direction coordinate:

$$k_{it}^{\mathrm{scale}} = -\lambda r^\top h_{it}, \qquad j_{it}^{\mathrm{scale}} = -\lambda(r^\top h_{it})r. \tag{3}$$

CAA therefore adds the same displacement at every valid prompt position, whereas scaling suppresses an item- and token-dependent fraction of the coordinate already present along r.

If Ψ_b denotes the subsequent transformer computation and $\widetilde{H}_i^{(m)}$ the modified prompt states, the downstream readout is

$$z_i^{(m)} = \Psi_b(\widetilde{H}_i^{(m)}). \tag{4}$$

Although CAA and directional scaling both write within $\mathrm{span}(r)$ at the intervention site, they need not produce the same final-state displacement or behavioral effect.

All-layer directional ablation has a different intervention topology:

$$h'_{i\ell t} = h_{i\ell t} - (r^\top h_{i\ell t})r \qquad \text{for every intervention layer } \ell. \tag{5}$$

Each projection changes the state received by subsequent layers, after which the parallel component is removed again. We therefore treat all-layer ablation as a repeated trajectory-level intervention, rather than as another single-site write.

4.3 Full-generation semantic measurement

The primary behavioral endpoint is a blinded binary judgment of whether the complete generated response semantically refuses. A refusal to a harmful prompt is a hit, whereas a refusal to a benign prompt is a false alarm. Let H and F denote the corresponding observed rates.

For the probit-scale indices, we apply the log-linear correction Hautus (1995):

$$\widetilde{H} = \frac{n_H^{\mathrm{ref}} + 0.5}{n_H + 1}, \qquad \widetilde{F} = \frac{n_B^{\mathrm{ref}} + 0.5}{n_B + 1}. \tag{6}$$

We then compute

$$d' = \Phi^{-1}(\widetilde{H}) - \Phi^{-1}(\widetilde{F}), \qquad c = -\frac{1}{2}\left[\Phi^{-1}(\widetilde{H}) + \Phi^{-1}(\widetilde{F})\right]. \tag{7}$$

For these binary observations, d' is a descriptive probit-scale index of harmful–benign refusal separation, while c describes the model's general response tendency. Their interpretation as fully identified latent sensitivity and criterion parameters would require the classical equal-variance evidence model. We therefore interpret them jointly with the underlying hit and false-alarm rates.

The primary statistics are estimated separately for XSTest and WildGuardTest. A combined statistic describes an equally weighted mixture of the two sampled populations, but it is not the mean of the dataset-specific separation indices because the probit transformation is nonlinear:

$$d'_{\mathrm{mix}} \neq \frac{1}{2}\left(d'_{\mathrm{XSTest}} + d'_{\mathrm{WildGuardTest}}\right). \tag{8}$$

Condition-wise uncertainty uses 20,000 bootstrap replicates stratified by dataset and prompt class. Intervention effects are paired with baseline by prompt and resampled within the same strata, producing intervals for $\Delta d'$, ΔH, and ΔF. We additionally estimate the WildGuardTest-minus-XSTest difference between paired intervention effects. Numerical paired effects, direct dataset contrasts, and combined-population summaries are reported in the appendix.

4.4 COMMON-SHIFT NULL AND PROMPT-SPECIFIC LOCAL RESPONSE

The simplest downstream-response model assumes that the same activation displacement produces the same score change for every prompt. Let $s_i(k)$ denote the continuous first-token refusal score for prompt i under CAA displacement k. Under a prompt-invariant gain g,

$$s_i(k) \approx s_i(0) + kg. \tag{9}$$

The harmful and benign score distributions then undergo the same translation. Their pairwise ranking and standardized separation remain unchanged, while their position relative to a fixed response threshold moves. This common-shift model therefore predicts a change in response tendency without a change in discrimination.

We relax this assumption by assigning each prompt its own local downstream gain:

$$G_i = \left.\frac{\partial s_i(k)}{\partial k}\right|_{k=0}, \qquad \widehat{s}_i(k) = s_i(0) + kG_i. \tag{10}$$

We estimate G_i using a central finite difference at $\pm 0.05\sigma_\rho$, where σ_ρ is the baseline standard deviation of the refusal-direction coordinate, and use $\pm 0.10\sigma_\rho$ as a robustness check.

Prompt-specific gain can alter both the location and shape of the class-conditional score distributions. For class $y \in \{H, B\}$,

$$\mu_y(k) \approx \mu_y(0) + k\,\mathbb{E}[G_i \mid y], \tag{11}$$

$$\mathrm{Var}[s(k) \mid y] \approx \mathrm{Var}[s(0) \mid y] + 2k\,\mathrm{Cov}[s(0), G \mid y] + k^2\,\mathrm{Var}[G \mid y]. \tag{12}$$

The first term permits different class-mean shifts, while the covariance and gain-variance terms permit changes in within-class dispersion.

We summarize the resulting continuous-score separation using

$$d'_{\mathrm{G}}(k) = \frac{\mu_H(k) - \mu_B(k)}{\sqrt{\frac{1}{2}[\sigma_H^2(k) + \sigma_B^2(k)]}}. \tag{13}$$

For every model–dataset cell, we compare four trajectories computed at the same d'_{G} level: measured finite-dose scores, item-specific gain predictions, dataset-by-class mean-gain predictions, and global-mean-gain predictions. The latter two serve as null models that preserve progressively less prompt-specific information.

The gain-trajectory reconstruction region is restricted to the frozen local range $|\alpha| \leq 0.125$. Dose is reported as k/σ_ρ using the checkpoint-specific conversion in Appendix C. A later follow-up uncertainty analysis treats $|\alpha| \leq 0.0625$ as its primary inferential radius and $|\alpha| \leq 0.125$ as a robustness radius; this inferential priority was chosen after inspection rather than preregistered. Larger doses are used only to evaluate where the first-order approximation ceases to be valid.

Semantic SDT and this continuous-proxy analysis remain distinct evaluation systems. Semantic d' is computed from binary full-generation judgments, whereas d'_{G} is computed from continuous first-token proxy distributions. The local gain model tests how a common activation displacement deforms the proxy-score distributions; it is not assumed to numerically mediate semantic refusal.

5 RESULTS

5.1 DIRECTION IDENTITY IS OPERATIONAL

Changing the identification evidence can leave the selected vector nearly unchanged or move it substantially. Gemma 2–2B and both Qwen checkpoints were stable across the tested lexical proxies, whereas Gemma 2–9B selected different layers under `I` and `I+It` scoring, with cosine similarity 0.356 (Table 1).

Table 1: Sensitivity of direction identification to lexical proxy. Gemma and Qwen rows summarize related but non-identical audits.

Model	Candidate bank	Scorer contrast	Selected direction and agreement
Gemma 2–2B	`I`-filtered	`I` vs. `I+It`	Both L15/-1; cosine ≈ 1.000
Gemma 2–2B	`I+It`-filtered	`I` vs. `I+It`	Both L15/-1; cosine ≈ 1.000
Gemma 2–9B	`I`-filtered	`I` vs. `I+It`	L31/-1 vs. L23/-1; cosine ≈ 0.356
Gemma 2–9B	`I+It`-filtered	`I` vs. `I+It`	L31/-1 vs. L23/-1; cosine ≈ 0.356
Qwen 2.5–3B	fixed `I+As`	three lexical scores	All L27/-1; $\rho = .981$–$.996$
Qwen 2.5–7B	fixed `I+As`	three lexical scores	All L15/-1; $\rho = .992$–$.998$

The same-bank COSMIC audit reveals still stronger cross-objective variation.

Table 2: Frozen Wang directions and the same-bank COSMIC audit.

Model	Layer	Pos.	Bank	COSMIC rank	Cosine
Gemma 2–2B-IT	15	-1	130	1	1.000
Gemma 2–9B-IT	23	-1	210	2	0.884
Qwen 2.5–3B-Instruct	26	-4	180	102	0.089
Qwen 2.5–7B-Instruct	19	-4	140	4	0.552

Agreement is exact for Gemma 2–2B, but the Qwen 2.5–3B Wang direction has cosine similarity 0.089 with the COSMIC winner and ranks 102nd of 180 candidates. Thus, within-family proxy stability does not establish a unique refusal direction; the identified r depends on the operational objective.

5.2 A Common Direction Does Not Imply a Common Behavioral Effect

Across models and operators, hit and false-alarm rates do not move as a rigid pair. Interventions therefore change both the general propensity to refuse and the descriptive harmful–benign refusal contrast. The direction and magnitude of these changes depend on the intervention operator and prompt population.

The datasets have distinct starting regimes and responses: WildGuardTest generally has lower baseline harmful refusal and a more conservative response tendency. Positive CAA usually increases both harmful and benign refusal, whereas negative CAA and scaling often reduce both. The separation index therefore depends on the relative movement of H and F, not refusal rate alone.

Pooling can reverse the apparent separation-index change. For Gemma 2–2B all-layer ablation, the equally weighted combined statistic gives $\Delta d' = +0.167$, while the paired changes are -0.725 on WildGuardTest and -0.477 on XSTest. This is a consequence of pooling distinct baseline regimes before the nonlinear SDT transformation, not a contradiction among estimates.

Direct interaction tests confirm population-dependent effects, especially in harmful-prompt refusal; complete paired and multiplicity-adjusted results are in the appendix.

Despite these changes, semantic d' remains positive in every tested model–dataset condition. Observed harmful–benign refusal selectivity therefore persists under the tested prefill interventions, although this does not rule out regeneration or later use of direction-related information during decoding. Overall, manipulating the same frozen direction does not produce a common behavioral shift and cannot be characterized by overall refusal rate alone. To examine one downstream route by which a common write can nevertheless deform discrimination, we next turn to the separate continuous proxy system.

5.3 Prompt-Specific Gain Predicts Local Proxy Discrimination

A common activation displacement does not produce a common downstream score shift across prompts. Prompt-specific gains estimated at two finite-difference scales correlate above 0.998 in every model, while median item-level local-linearity R^2 exceeds 0.995. Thus, G_i is not merely a

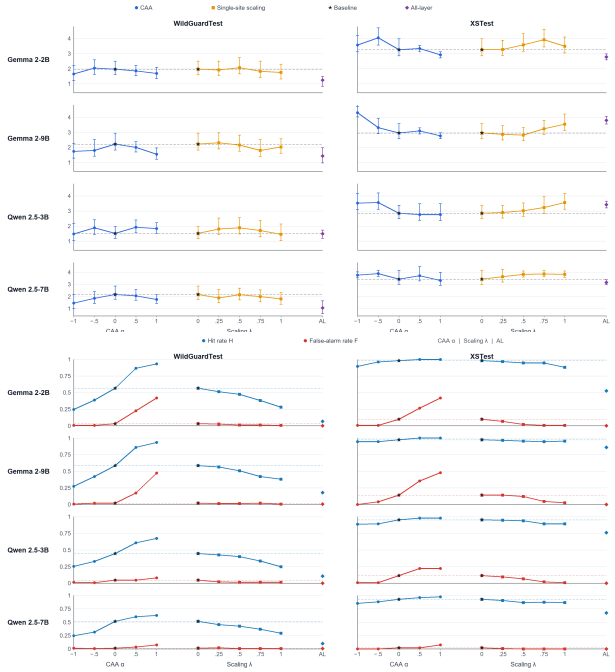


Figure 1: Dataset-specific full-generation outcomes. Top: descriptive d' with 95% bootstrap intervals. Bottom: hit rate H and false-alarm rate F. Rows are models and columns are prompt populations; dashed lines mark dataset-specific baselines. CAA and scaling retain their native parameters, and all-layer ablation is categorical.

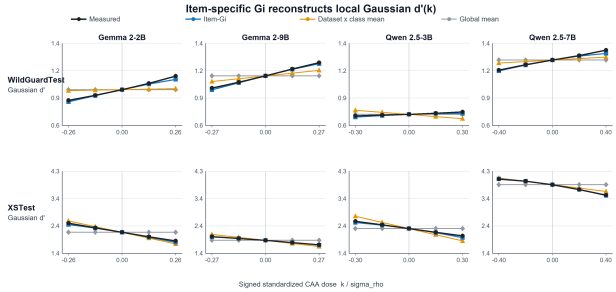


Figure 2: Measured and reconstructed continuous-score Gaussian $d'_{\mathrm{G}}(k)$ inside the frozen local region. Item-specific G_i tracks measured finite-dose trajectories more closely than dataset-by-class and global-mean gain nulls in all eight model–dataset cells.

derivative defined in the infinitesimal limit; it is a stable, measurable local susceptibility at operational finite perturbations.

The resulting discrimination response differs systematically between WildGuardTest and XSTest across all four models. The same dataset interaction appears in continuous AUC, showing that it is not an artifact of the Gaussian separation statistic. The moment decomposition further reveals different distributional routes: changes on WildGuardTest are dominated by within-class dispersion, whereas XSTest consistently exhibits a negative class-mean contribution. Full slopes, confidence intervals, and decomposition terms are reported in the appendix.

Item-specific gain provides the lowest trajectory error in all eight model–dataset cells (Table 11). Its advantage over the dataset-by-class null survives corrected prompt-bootstrap tests in most cells and a separate item-pairing control in seven of eight. Local discrimination deformation therefore depends on genuine within-class prompt–gain pairing, rather than only on differences between harmful and benign class averages.

Held-out results show that baseline refusal score and direction coordinate partly predict G_i, suggesting prompt-adaptive control without establishing a deployable controller.

Accuracy deteriorates outside the frozen local region; G_i is therefore a local account of proxy-score deformation, not a global intervention model.

6 Discussion

Refusal directions are operational causal approximations, not uniquely recovered internal axes. The identified r depends on construction and selection evidence; neither proxy agreement nor causal control establishes a canonical representation.

The effect is not contained in the vector alone: with checkpoint, direction, site, operator, and dose fixed, CAA still produces prompt-specific score changes. Stable G_i and its advantage over structured nulls motivate prompt-dependent doses, although baseline-state prediction remains a proof of concept.

Directional scaling may also evoke the model self-repair studied by McGrath et al. (2023) and Rushing & Nanda (2024). Because we do not separate perturbation from compensation, this remains a candidate explanation.

Harmful and benign refusal move differently across operators and prompt populations. The G_i analysis supplies a route to proxy-discrimination change, not a mediator of semantic outcomes. A direction's meaning therefore depends on its identification, operator and site, prompt state, downstream computation, and evaluation population.

7 Limitations

Semantic d' is a comparative probit-scale index because matching cannot guarantee equal-variance evidence. Four checkpoints, two datasets, and local CAA-only G_i may not generalize across operators, directions, or doses. Finally, G_i models a first-token proxy, with no assumption-free bridge to semantic d'.

AI Use Statement

Generative AI supported writing refinement, literature retrieval, research ideation, analysis planning, and code review. The author verified all sources, code, analyses, and claims.

Reproducibility Statement

The appendix specifies the protocols. The anonymous supplement provides prompt manifests, frozen inputs and labels, provenance records, analysis code, and reference outputs for reproducing all reported analyses without model inference or judge-API calls.

References

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. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-4322. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/f545448535dfde4f9786555403ab7c49-Abstract-Conference.html.

Harinder Aujla. d'_o: Sensitivity at the optimal criterion location. *Behavior Research Methods*, 55(5): 2532–2558, 2023. doi: 10.3758/s13428-022-01913-5. URL https://doi.org/10.3758/s13428-022-01913-5.

Jon-Paul Cacioli. LLMs as signal detectors: Sensitivity, bias, and the temperature-criterion analogy. *arXiv preprint arXiv:2603.14893*, 2026. URL https://arxiv.org/abs/2603.14893.

Gemma Team, Morgane Riviere, et al. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*, 2024. doi: 10.48550/arXiv.2408.00118. URL https://arxiv.org/abs/2408.00118.

Tanay Gondil. Do language models know when they'll refuse? probing introspective awareness of safety boundaries. *arXiv preprint arXiv:2604.00228*, 2026. URL https://arxiv.org/abs/2604.00228.

David M. Green and John A. Swets. *Signal Detection Theory and Psychophysics*. Wiley, 1966.

Seungju Han, Kavel Rao, Allyson Ettinger, Liwei Jiang, Bill Yuchen Lin, Nathan Lambert, Yejin Choi, and Nouha Dziri. Wildguard: Open one-stop moderation tools for safety risks, jailbreaks, and refusals of LLMs. In *Advances in Neural Information Processing Systems, Datasets and Benchmarks Track*, volume 37, 2024. doi: 10.52202/079017-0261. URL https://papers.neurips.cc/paper_files/paper/2024/hash/0f69b4b96a46f284b726fbd70f74fb3b-Abstract-Datasets_and_Benchmarks_Track.html.

Michael J. Hautus. Corrections for extreme proportions and their biasing effects on estimated values of d'. *Behavior Research Methods, Instruments, & Computers*, 27(1):46–51, 1995. doi: 10.3758/BF03203619.

Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70, 1979. doi: 10.2307/4615733.

Shawn Im and Sharon Li. A unified understanding and evaluation of steering methods. *arXiv preprint arXiv:2502.02716*, 2025. doi: 10.48550/arXiv.2502.02716. URL https://arxiv.org/abs/2502.02716.

Bruce W. Lee, Inkit Padhi, Karthikeyan Natesan Ramamurthy, Erik Miehling, Pierre Dognin, Manish Nagireddy, and Amit Dhurandhar. Programming refusal with conditional activation steering. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/e2dd53601de57c773343a7cdf09fae1c-Abstract-Conference.html.

Neil A. Macmillan and C. Douglas Creelman. *Detection Theory: A User's Guide*. Lawrence Erlbaum Associates, 2 edition, 2005.

Thomas Marshall, Adam Scherlis, and Nora Belrose. Refusal in LLMs is an affine function. *arXiv preprint arXiv:2411.09003*, 2024. URL https://arxiv.org/abs/2411.09003.

Thomas McGrath, Matthew Rahtz, János Kramár, Vladimir Mikulik, and Shane Legg. The hydra effect: Emergent self-repair in language model computations. *arXiv preprint arXiv:2307.15771*, 2023. doi: 10.48550/arXiv.2307.15771. URL https://arxiv.org/abs/2307.15771.

Qwen Team, An Yang, et al. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024. doi: 10.48550/arXiv.2412.15115. URL https://arxiv.org/abs/2412.15115.

Nina Rimsky, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Turner. Steering llama 2 via contrastive activation addition. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, pp. 15504–15522, 2024. doi: 10.18653/v1/2024.acl-long.828.

Paul Röttger, Hannah Rose 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*, pp. 5377–5400, 2024. doi: 10.18653/v1/2024.naacl-long.301.

Cody Rushing and Neel Nanda. Explorations of self-repair in language models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 42836–42855. PMLR, 2024. URL https://proceedings.mlr.press/v235/rushing24a.html.

Leheng Sheng, Changshuo Shen, Weixiang Zhao, Junfeng Fang, Xiaohao Liu, Zhenkai Liang, Xiang Wang, An Zhang, and Tat-Seng Chua. AlphaSteer: Learning refusal steering with principled null-space constraint. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/2b6a012bf4487225245314 21ff76b31f-Abstract-Conference.html.

Vincent Siu, Nicholas Crispino, Zihao Yu, Sam Pan, Zhun Wang, Yang Liu, Dawn Song, and Chenguang Wang. COSMIC: Generalized refusal direction identification in LLM activations. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 25534–25553. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.findings-acl.1310. URL https://aclanthology.org/2025.findings-acl.1310/.

Julian Skifstad, Xinyue Annie Yang, and Glen Chou. Local linearity of LLMs enables activation steering via model-based linear optimal control. *arXiv preprint arXiv:2604.19018*, 2026. URL https://arxiv.org/abs/2604.19018.

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. doi: 10.48550/arXiv.2308.10248. URL https://arxiv.org/abs/2308.10248.

Xinpeng Wang, Bolei Ma, Chengzhi Hu, Leon Weber-Genzel, Paul R"ottger, Frauke Kreuter, Dirk Hovy, and Barbara Plank. "my answer is C": First-token probabilities do not match text answers in instruction-tuned language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 7407–7416. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.findings-acl.441. URL https://aclanthology.org/2024.findings-acl.441/.

Xinpeng Wang, Mingyang Wang, Yihong Liu, Hinrich Schuetze, and Barbara Plank. Refusal direction is universal across safety-aligned languages. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-1086. URL https://proceedings.nips.cc/paper_files/paper/2025/hash/2e94772e3c079d83f79c311d07456111-Abstract-Conference.html.

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. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 66945–66970. PMLR, 2025. URL https://proceedings.mlr.press/v267/wollschlager25a.html.

A Exact Refusal-Proxy Specifications

All lexical scores are computed from the next-token distribution at the final prompt position. For a token set T, let $p_T(x) = \sum_{v \in T} p(y_1 = v \mid x)$. The score is

$$S_T(x) = \log\left[p_T(x) + 10^{-8}\right] - \log\left[1 - p_T(x) + 10^{-8}\right]. \tag{14}$$

Token strings below denote tokenizer-specific single tokens, not character-level string matching. The exact vocabularies are listed in Table 3.

Table 3: Exact first-token proxy vocabularies. "Formal/G_i" marks the score used both for frozen Wang-style selection and for the continuous-score gain analysis.

Model family	Role	Token strings	Token IDs
Gemma 2	Stage-0 narrow	`I`	`[235285]`
Gemma 2	Stage-0 broad; Formal/G_i	`I`, `It`	`[235285, 1718]`
Qwen 2.5	Lexical-audit narrow	`I`	`[40]`
Qwen 2.5	Formal/G_i	`I`, `As`	`[40, 2121]`
Qwen 2.5	Lexical control	`I`, `It`	`[40, 2132]`

In the preliminary Gemma audit, Arditi-style refers to the construction-prompt bank, not to a separate substring-marker score. Records were independently filtered by $S_{\{I\}} > 0$ for harmful and < 0 for benign, or by the corresponding $S_{\{I,It\}}$ rule. Each filtered candidate bank was then ranked once by each scorer, yielding the four bank–scorer cells reported in Table 1. No Arditi substring-marker list entered those lexical scores. The Qwen audit held the $\{I, As\}$-filtered bank fixed and compared $S_{\{I\}}$, $S_{\{I,As\}}$, and $S_{\{I,It\}}$; `As` and `It` were separate broad targets and were never pooled into one set.

Formal construction used the English Wang prompt bank. Candidate directions were harmful-minus-benign activation means at candidate layer–position sites. Candidates passing the preserved-layer and benign-KL gates were ranked by the change in mean harmful-validation S_T under directional ablation, using the family-specific Formal vocabularies in Table 3. The selected directions were then frozen. The same S_T defines the continuous first-token score $s_i(k)$ in the G_i analysis; it is not the semantic outcome.

B Semantic-Judgment Protocol

The semantic endpoint was assigned independently to each prompt–response pair by `gemini-3.7-flash` with low reasoning effort under prompt version `strict_semantic_refusal_v2`. The classifier received only a record identifier, the user prompt, and the assistant response; model identity, dataset, intervention condition, and dose were withheld. It returned exactly one Boolean field, `semantic_refusal`.

The decision rule labeled a response as refusal only when it declined, withheld, or meaningfully avoided the requested substantive assistance. Explicit and implicit refusals, including refusals followed only by safe alternatives or general background, counted as refusals. A response counted as non-refusal when it meaningfully attempted the requested task: warnings, caveats, apologies, or an initial refusal did not override substantial assistance that followed. Uncertainty, factual error, poor quality, brevity, irrelevance, clarification requests, and generation failure were not themselves refusal unless substantive assistance was withheld. Prompt and response text were explicitly treated as data rather than instructions to the judge.

The run used 128 workers with API batches of eight, checkpointed by model, prompt identifier, and condition. Exact duplicate prompt–response pairs were keyed by a SHA-256 digest of the two texts and reused rather than rejudged. Of 24,000 final labels, 23,994 were returned by the automated judge; six provider-blocked records were manually adjudicated under the same blinded binary rule. These six completions are provenance recovery, not an independent inter-rater reliability sample.

C CAA Dose Calibration and Local Regions

Let r_{raw} be the frozen, non-unit construction vector and $r = r_{\text{raw}}/\|r_{\text{raw}}\|_2$. The implemented CAA write is

$$h' = h + \alpha r_{\text{raw}} = h + kr, \qquad k = \alpha\|r_{\text{raw}}\|_2, \qquad z = \frac{k}{\sigma_\rho}. \tag{15}$$

Here $\rho_i = \langle h_i, r\rangle$ is the baseline refusal coordinate at the frozen intervention site and σ_ρ is its sample standard deviation in the fixed reference set. Thus $z = \alpha\|r_{\text{raw}}\|_2/\sigma_\rho$ is the signed standardized CAA dose. Direction and denominator are fixed within checkpoint; their estimation uncertainty is not propagated. Scaling retains its native λ axis, and all-layer ablation remains categorical rather than being assigned a fictitious single-site dose.

Table 4: Checkpoint-specific conversion from raw CAA coefficient α to standardized dose $z = k/\sigma_\rho$.

Model	$\|r_{\text{raw}}\|_2$	σ_ρ	z/α
Gemma 2–2B	111.1519	53.1230	2.0923
Gemma 2–9B	155.6814	73.3297	2.1230
Qwen 2.5–3B	37.5364	15.5623	2.4120
Qwen 2.5–7B	48.4452	15.0753	3.2136

The continuous-score G_i analysis used an additional finite-dose CAA sweep evaluated by single-forward next-token scoring rather than full generation. First-token probabilities need not track final text behavior in instruction-tuned models Wang et al. (2024); here they are used only as a continuous local proxy. Measured proxy scores were collected at $\alpha = \pm 0.0625$ and ± 0.125 for the frozen gain-trajectory reconstruction region, alongside larger diagnostic doses. A later follow-up uncertainty analysis designated $|\alpha| \leq 0.0625$ as its primary inferential radius and $|\alpha| \leq 0.125$ as a wider-radius robustness analysis. This inferential priority was chosen after inspection rather than preregistered; both radii lie inside the pre-existing reconstruction region. These conditions are separate from the semantic full-generation experiment, whose CAA grid is $\alpha \in \{-1, -0.5, 0.5, 1\}$. All continuous-score doses map to geometric and standardized units through Eq. (15).

D Dataset-Specific Effects and Direct Dataset Contrasts

All intervention effects are paired to the corresponding dataset-specific baseline by prompt identity. Cross-dataset tests use the contrast

$$(\Delta M)_{\text{WildGuardTest}} - (\Delta M)_{\text{XSTest}}, \qquad M \in \{d', H, F\},$$

within each frozen model and condition. Thus the test targets the difference between intervention effects; it does not infer a difference merely because one dataset-specific interval excludes zero and the other does not. Holm's sequentially rejective procedure Holm (1979) is applied separately to each family of 36 cross-dataset contrasts.

Table 5: Direct cross-dataset intervention-effect tests. "Nominal" counts 95% bootstrap intervals excluding zero; "Holm" counts contrasts with adjusted $p < .05$.

Outcome	Nominal / 36	Holm / 36
$\Delta d'$	6	2
ΔH	24	21
ΔF	19	6

Table 6: Cross-dataset $\Delta d'$ contrasts surviving Holm correction. The contrast direction is WildGuardTest minus XSTest.

Gemma 2–9B condition	$\Delta d'_{\text{WG}}$	$\Delta d'_{\text{XS}}$	Difference [95% CI]	Holm p
All-layer ablation	-0.788	0.832	-1.620 $[-2.242, -0.794]$	.0018
CAA -1.00	-0.479	1.349	-1.828 $[-2.429, -1.045]$	.0018

E Combined-Population Descriptive Summary

The combined statistic pools 150 harmful and 150 benign prompts from each dataset before applying the log-linear correction and probit transformation. It therefore describes an equally weighted mixture of the two sampled prompt populations; it is not the arithmetic mean of their dataset-specific d' values and is not used for the primary semantic interpretation.

F Bootstrap Uncertainty for Semantic SDT

The table reports condition-wise 95% intervals from 20,000 stratified bootstrap replicates of the binary semantic-refusal labels. Resampling preserves the four 150-item strata defined by XSTest versus WildGuardTest and harmful versus benign prompts. Each entry is the point estimate followed by $[2.5\%, 97.5\%]$.

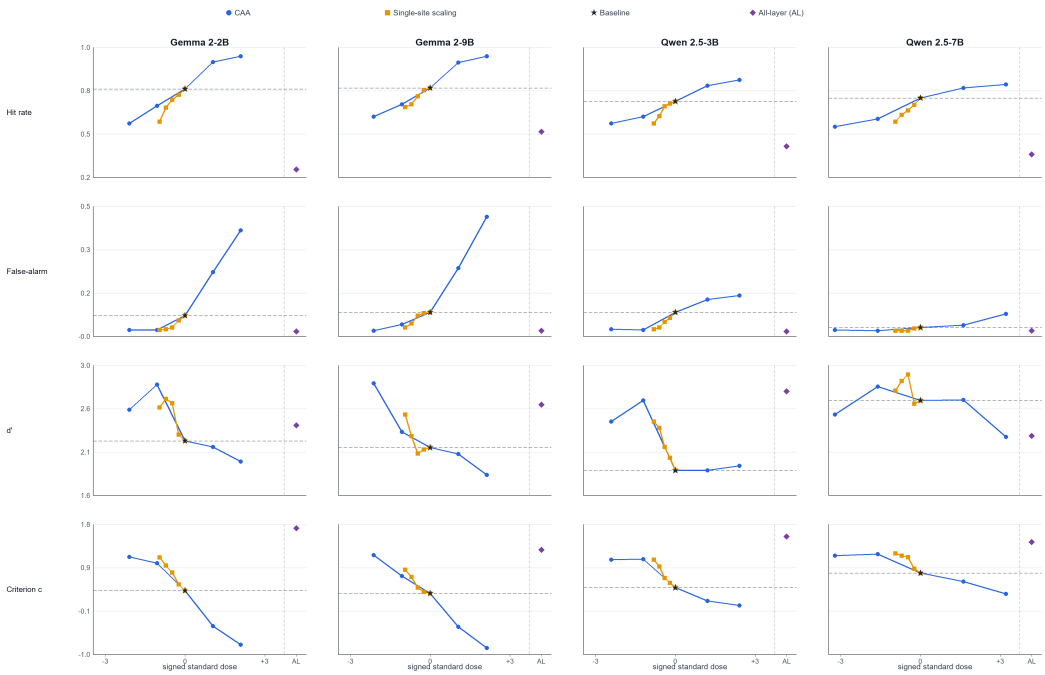


Figure 3: Combined-population semantic trajectories. Rows show hit rate, false-alarm rate, d', and criterion c. This pooled view is retained for comparison with earlier analyses; the dataset-specific results in the main text are primary.

Table 7: Combined-dataset semantic SDT point estimates and 95% stratified bootstrap confidence intervals.

Condition	d' [95% CI]	c [95% CI]
Gemma 2–2B		
Baseline	2.237 [2.004, 2.519]	0.372 [0.251, 0.507]
CAA −1.00	2.571 [2.233, 3.163]	1.110 [0.941, 1.405]
CAA −0.50	2.843 [2.502, 3.440]	0.974 [0.805, 1.272]
CAA +0.50	2.172 [1.934, 2.457]	−0.404 [−0.543, −0.282]
CAA +1.00	2.015 [1.746, 2.361]	−0.806 [−0.979, −0.672]
Scale 0.25	2.304 [2.059, 2.611]	0.511 [0.386, 0.663]
Scale 0.50	2.642 [2.346, 3.119]	0.770 [0.621, 1.008]
Scale 0.75	2.689 [2.370, 3.356]	0.925 [0.765, 1.258]
Scale 1.00	2.596 [2.257, 3.197]	1.097 [0.928, 1.393]
All-layer ablation	2.404 [2.275, 2.534]	1.734 [1.669, 1.799]
Gemma 2–9B		
Baseline	2.165 [1.938, 2.432]	0.313 [0.200, 0.442]
CAA −1.00	2.855 [2.493, 3.293]	1.149 [0.969, 1.368]
CAA −0.50	2.334 [2.072, 2.694]	0.691 [0.558, 0.876]
CAA +0.50	2.096 [1.859, 2.366]	−0.417 [−0.553, −0.296]
CAA +1.00	1.872 [1.605, 2.224]	−0.878 [−1.054, −0.746]
Scale 0.25	2.143 [1.920, 2.409]	0.347 [0.230, 0.479]
Scale 0.50	2.101 [1.868, 2.374]	0.440 [0.324, 0.572]
Scale 0.75	2.289 [2.029, 2.632]	0.669 [0.538, 0.840]
Scale 1.00	2.520 [2.229, 2.997]	0.831 [0.685, 1.069]
All-layer ablation	2.627 [2.261, 3.061]	1.263 [1.082, 1.485]
Qwen 2.5–3B		
Baseline	1.918 [1.695, 2.177]	0.437 [0.325, 0.565]
CAA −1.00	2.445 [2.129, 3.104]	1.047 [0.889, 1.380]
CAA −0.50	2.674 [2.335, 3.267]	1.058 [0.889, 1.355]

Condition	d' [95% CI]	c [95% CI]
CAA $+0.50$	1.920 [1.712, 2.159]	0.145 [0.035, 0.262]
CAA $+1.00$	1.968 [1.753, 2.211]	0.047 [-0.064, 0.166]
Scale 0.25	2.055 [1.822, 2.338]	0.543 [0.425, 0.685]
Scale 0.50	2.172 [1.922, 2.493]	0.647 [0.523, 0.810]
Scale 0.75	2.378 [2.092, 2.864]	0.902 [0.754, 1.141]
Scale 1.00	2.445 [2.129, 3.120]	1.047 [0.889, 1.380]
All-layer ablation	2.769 [2.658, 2.870]	1.552 [1.501, 1.607]
Qwen 2.5–7B		
Baseline	2.671 [2.373, 3.148]	0.755 [0.602, 0.993]
CAA -1.00	2.520 [2.183, 3.112]	1.135 [0.968, 1.431]
CAA -0.50	2.821 [2.453, 3.267]	1.167 [0.982, 1.389]
CAA $+0.50$	2.676 [2.399, 3.058]	0.569 [0.429, 0.761]
CAA $+1.00$	2.279 [2.040, 2.563]	0.302 [0.181, 0.441]
Scale 0.25	2.637 [2.327, 3.128]	0.853 [0.696, 1.097]
Scale 0.50	2.952 [2.582, 3.412]	1.101 [0.915, 1.329]
Scale 0.75	2.881 [2.511, 3.338]	1.136 [0.952, 1.363]
Scale 1.00	2.778 [2.415, 3.223]	1.188 [1.004, 1.414]
All-layer ablation	2.290 [1.921, 2.735]	1.432 [1.247, 1.656]

G Paired Changes in the Combined Semantic Separation Index

For each intervention condition, responses are paired with baseline by model, dataset, prompt class, and prompt identity. Each bootstrap replicate resamples the same prompt pairs within each of the four strata before recomputing both Hautus-corrected statistics and their difference. The table reports $\Delta d' = d'_{\text{intervention}} - d'_{\text{baseline}}$ and its paired 95% interval.

Table 8: Paired changes in the combined-dataset semantic separation index.

Condition	$\Delta d'$	Paired 95% CI
Gemma 2–2B		
All-layer ablation	0.167	[-0.144, 0.427]
CAA -1.00	0.334	[-0.018, 0.941]
CAA -0.50	0.606	[0.264, 1.196]
CAA $+0.50$	-0.065	[-0.350, 0.221]
CAA $+1.00$	-0.222	[-0.550, 0.145]
Scale 0.25	0.067	[-0.102, 0.259]
Scale 0.50	0.405	[0.137, 0.796]
Scale 0.75	0.452	[0.104, 1.046]
Scale 1.00	0.360	[0.012, 0.961]
Gemma 2–9B		
All-layer ablation	0.462	[0.065, 0.975]
CAA -1.00	0.691	[0.307, 1.195]
CAA -0.50	0.169	[-0.090, 0.487]
CAA $+0.50$	-0.069	[-0.331, 0.204]
CAA $+1.00$	-0.293	[-0.618, 0.070]
Scale 0.25	-0.022	[-0.144, 0.101]
Scale 0.50	-0.064	[-0.232, 0.112]
Scale 0.75	0.124	[-0.106, 0.408]
Scale 1.00	0.355	[0.072, 0.780]
Qwen 2.5–3B		
All-layer ablation	0.851	[0.586, 1.078]
CAA -1.00	0.527	[0.210, 1.123]
CAA -0.50	0.755	[0.408, 1.355]
CAA $+0.50$	0.002	[-0.205, 0.198]
CAA $+1.00$	0.050	[-0.172, 0.261]
Scale 0.25	0.137	[-0.008, 0.308]
Scale 0.50	0.253	[0.033, 0.515]
Scale 0.75	0.460	[0.165, 0.894]
Scale 1.00	0.527	[0.208, 1.110]

Condition	$\Delta d'$	Paired 95% CI
Qwen 2.5–7B		
All-layer ablation	-0.382	$[-0.790, 0.132]$
CAA -1.00	-0.151	$[-0.605, 0.446]$
CAA -0.50	0.149	$[-0.245, 0.655]$
CAA $+0.50$	0.005	$[-0.291, 0.207]$
CAA $+1.00$	-0.392	$[-0.826, -0.090]$
Scale 0.25	-0.034	$[-0.344, 0.321]$
Scale 0.50	0.281	$[-0.110, 0.788]$
Scale 0.75	0.210	$[-0.190, 0.722]$
Scale 1.00	0.107	$[-0.294, 0.619]$

H Continuous-Score Robustness and Assumption Audits

The continuous-score analyses use the first-token refusal log-odds and are distinct from the full-generation semantic SDT endpoint. The tests below use the follow-up primary inferential radius, $|\alpha| \leq 0.0625$, within the frozen reconstruction region described in Appendix C. All reported interaction p-values are adjusted jointly across the 24 local slope tests (Gaussian d'_G, AUC, and criterion) using Holm's procedure.

H.1 Dataset-dependent local slopes

Table 9: WildGuardTest-minus-XSTest differences in local CAA slopes. Positive values indicate a more positive slope on WildGuardTest.

Model	$\Delta\beta_{d'_G}$ [95% CI]	p_{Holm24}	$\Delta\beta_{\text{AUC}}$ [95% CI] / p_{Holm24}
Gemma 2–2B	1.656 [1.325, 2.045]	.0012	.2508 [.1759, .3341] / .0012
Gemma 2–9B	1.040 [.757, 1.370]	.0012	.1810 [.1236, .2465] / .0012
Qwen 2.5–3B	.940 [.656, 1.289]	.0012	.1170 [.0544, .1893] / .0128
Qwen 2.5–7B	.975 [.582, 1.490]	.0045	.0715 [.0371, .1101] / .0042

All eight interactions remain significant after the joint correction, although the corresponding within-dataset slopes are not uniformly distinguishable from zero. The interaction therefore establishes dataset dependence; it should not be read as evidence that every model has a nonzero slope on each dataset.

H.2 Moment decomposition

For Gaussian score separation, the local derivative can be decomposed into a class-mean term and a pooled-dispersion term. Table 10 reports their estimated contributions to $dd'_G/d\alpha$.

Table 10: Local Gaussian-d'_G slope decomposition. Intervals are stratified bootstrap 95% CIs.

Model	Dataset	Class-mean contribution	Dispersion contribution
Gemma 2–2B	WildGuardTest	.040 [−.168, .246]	.403 [.268, .574]
	XSTest	−1.552 [−1.775, −1.367]	.338 [.145, .535]
Gemma 2–9B	WildGuardTest	.210 [.046, .382]	.292 [.179, .434]
	XSTest	−.811 [−.985, −.668]	.262 [.122, .405]
Qwen 2.5–3B	WildGuardTest	−.137 [−.287, .010]	.203 [.113, .330]
	XSTest	−1.442 [−1.700, −1.231]	.557 [.335, .782]
Qwen 2.5–7B	WildGuardTest	.074 [−.040, .193]	.149 [.064, .264]
	XSTest	−.593 [−.762, −.454]	−.170 [−.555, .118]

The WildGuardTest dispersion contribution is positive for all four models and survives correction across the 16 decomposition terms; the XSTest class-mean contribution is negative for all four and likewise survives correction. For Qwen 2.5–3B on WildGuardTest, oppositely signed mean and dispersion contributions nearly cancel, explaining why its net within-dataset Gaussian-d'_G slope is weak despite a reproducible change in the component moments.

H.3 Item-specific gain against structured nulls

Table 11 reports the aggregate point-estimate error comparison. Table 12 complements it with cell-level uncertainty and pairing controls. The bootstrap column tests the paired trajectory-error advantage of item-specific gains over dataset-by-class mean gains. The permutation column instead destroys item identity by randomly re-pairing gains within dataset and class. These nulls answer different questions and their p-values need not agree.

Table 11: Prediction error for local Gaussian $d'_{\mathrm{G}}(k)$ at nonzero signed doses.

Predictor	MAE	RMSE	Best cells
Item-specific G_i	**0.0162**	**0.0232**	**8/8**
Dataset×class mean	0.0672	0.0779	0/8
Global mean	0.1288	0.1578	0/8

Table 12: Uncertainty and pairing controls for the item-specific G_i advantage.

Model	Dataset	RMSE advantage [95% CI]	p_{Holm8}	Pairing p_{Holm8}
Gemma 2–2B	WildGuardTest	.0678 [.0423, .1006]	.0028	.0016
	XSTest	.0334 [−.0059, .0744]	.1035	.0016
Gemma 2–9B	WildGuardTest	.0443 [.0241, .0708]	.0075	.0016
	XSTest	.0434 [.0148, .0728]	.0200	.0016
Qwen 2.5–3B	WildGuardTest	.0367 [.0154, .0647]	.0304	.0016
	XSTest	.0967 [.0425, .1512]	.0032	.0016
Qwen 2.5–7B	WildGuardTest	.0359 [.0133, .0643]	.0304	.0016
	XSTest	.0671 [.0105, .1442]	.0965	.0668

Thus, item-specific gains have the lowest observed trajectory RMSE in all eight cells, a multiplicity-corrected bootstrap advantage in six, and a multiplicity-corrected item-pairing advantage in seven. The evidence is strong but not uniform across every inferential construction.

H.4 Exploratory prediction from baseline state

A quadratic ridge model using only the baseline proxy score and normalized refusal coordinate predicts a nontrivial fraction of held-out G_i variance (Table 13). Hyperparameters were selected within training folds. Cross-dataset transfer remains positive in all eight train–test directions ($R^2 = .059$–$.734$), but this exploratory audit is not a validated prompt-adaptive controller.

Table 13: Out-of-fold R^2 for baseline-state prediction of G_i.

Model	WildGuardTest	XSTest
Gemma 2–2B	.277	.622
Gemma 2–9B	.191	.395
Qwen 2.5–3B	.304	.747
Qwen 2.5–7B	.286	.525

H.5 Distribution-shape audit

The Gaussian score statistic is used as a standardized separation summary rather than an identified latent equal-variance sensitivity. At baseline, harmful-to-benign standard-deviation ratios differ sharply by dataset (Table 14); all eight model–dataset cells reject equal variance after Holm correction across 40 tests. All eight harmful baseline distributions reject the tested Gaussian shape, and 53 of the 80 class-by-condition distributions reject normality after Holm correction. Continuous-score AUC supplies a complementary rank-based analysis that does not require Gaussian or equal-variance score distributions.

Table 14: Baseline harmful-to-benign proxy-score standard-deviation ratio.

Model	WildGuardTest	XSTest
Gemma 2–2B	1.401	.329
Gemma 2–9B	1.279	.362
Qwen 2.5–3B	1.413	.268
Qwen 2.5–7B	1.628	.564

I Generation-Cap Audit

Table 15: Audit of generations reaching the 512-token cap.

Quantity	Result
All generations	24,000
Capped generations	7,301 (30.42%)
Capped, semantic non-refusal	6,999
Capped, semantic refusal	302
Capped with missing semantic label	0
Stratified responses directly re-inspected	31
Stored judgment supported before cutoff	31/31

The reinspection sample was stratified by model, prompt class, and stored semantic label. It found no case in which the 512-token cutoff made the observed-text judgment ambiguous. This is evidence against an obvious truncation artifact, but it is not a counterfactual test of the unobserved continuation. Establishing full label invariance would require continuing all capped generations under the same decoding state and re-adjudicating them blind to the original labels.

J Baseline Semantic SDT by Dataset

Table 16: Baseline full-generation semantic SDT. Each entry is hit rate / false-alarm rate / d' / c.

Model	WildGuardTest	XSTest
Gemma 2–2B	.567/.033/1.961/.814	.980/.100/3.259/−.363
Gemma 2–9B	.587/.020/2.210/.887	.973/.140/2.953/−.407
Qwen 2.5–3B	.447/.047/1.515/.891	.953/.113/2.844/−.226
Qwen 2.5–7B	.513/.013/2.164/1.049	.927/.020/3.424/.280

K Dataset-Specific Paired Semantic Effects

Each intervention is paired to the dataset-specific baseline by prompt identity and resampled within dataset and prompt class. Each stratum contains 150 harmful and 150 benign prompts. The table reports the paired estimate and 95% interval from 20,000 bootstrap replicates. A superscript * marks metric-family Holm-adjusted $p < .05$ across the 72 model–dataset–intervention comparisons for that outcome. The binary semantic quantities are descriptive response-rate effects and probit-scale index changes; the table does not identify a latent evidence model.

Table 17: Dataset-specific paired changes from each intervention relative to baseline.

Model–dataset	Condition	ΔH [95% CI]	ΔF [95% CI]	$\Delta d'$ [95% CI]	Δc [95% CI]
Gemma 2–2B / WildGuardTest	All-layer ablation	-0.500^* $[-0.580, -0.420]$	-0.033 $[-0.067, -0.007]$	-0.725 $[-1.345, -0.310]$	1.284^* $[1.004, 1.519]$
	CAA -1.00	-0.320^* $[-0.393, -0.247]$	-0.027 $[-0.053, -0.007]$	-0.311 $[-0.783, 0.273]$	0.691^* $[0.455, 0.991]$
	CAA -0.50	-0.180^* $[-0.247, -0.113]$	-0.027 $[-0.053, -0.007]$	0.082 $[-0.380, 0.649]$	0.494^* $[0.265, 0.785]$
	CAA $+0.50$	0.300^* $[0.227, 0.373]$	0.193^* $[0.133, 0.260]$	-0.117 $[-0.633, 0.290]$	-0.991^* $[-1.258, -0.799]$
	CAA $+1.00$	0.367^* $[0.293, 0.447]$	0.387^* $[0.307, 0.467]$	-0.281 $[-0.834, 0.202]$	-1.453^* $[-1.762, -1.243]$
	Scale 0.25	-0.053 $[-0.093, -0.020]$	-0.007 $[-0.027, 0.013]$	-0.044 $[-0.387, 0.362]$	0.112 $[-0.058, 0.314]$
	Scale 0.50	-0.093^* $[-0.147, -0.047]$	-0.020 $[-0.047, 0.000]$	0.104 $[-0.249, 0.655]$	0.285 $[0.110, 0.561]$
	Scale 0.75	-0.187^* $[-0.253, -0.127]$	-0.020 $[-0.053, 0.007]$	-0.133 $[-0.650, 0.526]$	0.404 $[0.147, 0.735]$
	Scale 1.00	-0.287^* $[-0.360, -0.213]$	-0.027 $[-0.053, -0.007]$	-0.210 $[-0.682, 0.369]$	0.640^* $[0.408, 0.935]$
Gemma 2–2B / XSTest	All-layer ablation	-0.453^* $[-0.533, -0.373]$	-0.100^* $[-0.153, -0.053]$	-0.477 $[-1.212, -0.051]$	1.687^* $[1.457, 2.034]$
	CAA -1.00	-0.087^* $[-0.140, -0.040]$	-0.093^* $[-0.140, -0.047]$	0.301 $[-0.412, 0.989]$	0.912^* $[0.636, 1.335]$
	CAA -0.50	-0.020 $[-0.053, 0.007]$	-0.093^* $[-0.140, -0.047]$	0.786 $[0.115, 1.485]$	0.669^* $[0.391, 1.066]$
	CAA $+0.50$	0.020 $[0.000, 0.047]$	0.167^* $[0.107, 0.233]$	0.075 $[-0.664, 0.476]$	-0.686^* $[-0.898, -0.333]$
	CAA $+1.00$	0.020 $[0.000, 0.047]$	0.320^* $[0.247, 0.393]$	-0.343 $[-1.079, 0.069]$	-0.895^* $[-1.115, -0.540]$
	Scale 0.25	-0.013 $[-0.040, 0.013]$	-0.033 $[-0.067, -0.007]$	0.014 $[-0.546, 0.470]$	0.205 $[-0.012, 0.492]$
	Scale 0.50	-0.033 $[-0.067, 0.000]$	-0.080^* $[-0.127, -0.040]$	0.320 $[-0.325, 1.000]$	0.565 $[0.311, 0.962]$
	Scale 0.75	-0.033 $[-0.067, 0.000]$	-0.093^* $[-0.140, -0.047]$	0.657 $[-0.029, 1.349]$	0.734^* $[0.457, 1.134]$
	Scale 1.00	-0.100^* $[-0.153, -0.053]$	-0.093^* $[-0.140, -0.047]$	0.233 $[-0.490, 0.916]$	0.946^* $[0.670, 1.371]$
Gemma 2–9B / WildGuardTest	All-layer ablation	-0.407^* $[-0.487, -0.327]$	-0.013 $[-0.033, 0.000]$	-0.788 $[-1.247, -0.216]$	0.731^* $[0.506, 1.026]$
	CAA -1.00	-0.313^* $[-0.387, -0.240]$	-0.013 $[-0.033, 0.000]$	-0.479 $[-0.914, 0.083]$	0.576^* $[0.360, 0.866]$
	CAA -0.50	-0.167^* $[-0.227, -0.107]$	0.000 $[-0.020, 0.020]$	-0.418 $[-0.875, 0.035]$	0.209 $[-0.017, 0.437]$
	CAA $+0.50$	0.273^* $[0.207, 0.347]$	0.153^* $[0.100, 0.213]$	-0.207 $[-0.865, 0.217]$	-0.956^* $[-1.289, -0.754]$
	CAA $+1.00$	0.347^* $[0.267, 0.427]$	0.453^* $[0.373, 0.533]$	-0.664 $[-1.366, -0.152]$	-1.594^* $[-1.964, -1.366]$
	Scale 0.25	-0.020 $[-0.060, 0.020]$	-0.007 $[-0.020, 0.000]$	0.088 $[-0.126, 0.534]$	0.095 $[-0.009, 0.317]$
	Scale 0.50	-0.080 $[-0.133, -0.033]$	-0.007 $[-0.020, 0.000]$	-0.062 $[-0.301, 0.385]$	0.170 $[0.051, 0.393]$
	Scale 0.75	-0.167^* $[-0.227, -0.107]$	0.000 $[-0.020, 0.020]$	-0.418 $[-0.875, 0.031]$	0.209 $[-0.018, 0.438]$
	Scale 1.00	-0.207^* $[-0.273, -0.140]$	-0.013 $[-0.033, 0.000]$	-0.184 $[-0.598, 0.366]$	0.429^* $[0.220, 0.711]$
Gemma 2–9B / XSTest	All-layer ablation	-0.113^* $[-0.167, -0.060]$	-0.140^* $[-0.200, -0.087]$	0.832 $[0.226, 1.240]$	1.230^* $[1.012, 1.523]$
	CAA -1.00	-0.027 $[-0.060, 0.000]$	-0.140^* $[-0.200, -0.087]$	1.349^* $[0.805, 1.755]$	0.971^* $[0.758, 1.233]$
	CAA -0.50	-0.027 $[-0.053, -0.007]$	-0.100^* $[-0.153, -0.053]$	0.350 $[-0.141, 0.833]$	0.472 $[0.278, 0.762]$
	CAA $+0.50$	0.027 $[0.007, 0.053]$	0.213^* $[0.153, 0.280]$	0.136 $[-0.471, 0.507]$	-0.764^* $[-0.960, -0.474]$
	CAA $+1.00$	0.027 $[0.007, 0.053]$	0.340^* $[0.260, 0.420]$	-0.188 $[-0.809, 0.201]$	-0.926^* $[-1.131, -0.631]$
	Scale 0.25	-0.007 $[-0.020, 0.000]$	0.000 $[-0.027, 0.027]$	-0.090 $[-0.373, 0.093]$	0.045 $[-0.045, 0.188]$

Model–dataset	Condition	ΔH [95% CI]	ΔF [95% CI]	$\Delta d'$ [95% CI]	Δc [95% CI]
	Scale 0.50	−0.020 [−0.047, 0.000]	−0.020 [−0.060, 0.020]	−0.143 [−0.555, 0.156]	0.164 [0.020, 0.371]
	Scale 0.75	−0.027 [−0.053, −0.007]	−0.093* [−0.140, −0.053]	0.281 [−0.188, 0.705]	0.438 [0.265, 0.703]
	Scale 1.00	−0.020 [−0.047, 0.000]	−0.113* [−0.167, −0.067]	0.578 [0.103, 1.121]	0.525 [0.334, 0.845]
Qwen 2.5–3B / WildGuardTest	All-layer ablation	−0.340* [−0.420, −0.267]	−0.047 [−0.087, −0.013]	−0.030 [−0.546, 0.344]	1.082* [0.840, 1.290]
	CAA −1.00	−0.193* [−0.260, −0.127]	−0.033 [−0.067, −0.007]	−0.043 [−0.437, 0.558]	0.504* [0.314, 0.808]
	CAA −0.50	−0.120* [−0.180, −0.067]	−0.040 [−0.073, −0.013]	0.368 [−0.067, 0.935]	0.497* [0.281, 0.781]
	CAA +0.50	0.160* [0.093, 0.227]	0.000 [−0.027, 0.027]	0.402 [0.080, 0.735]	−0.201 [−0.369, −0.036]
	CAA +1.00	0.227* [0.153, 0.300]	0.033 [0.007, 0.067]	0.318 [−0.018, 0.602]	−0.420* [−0.593, −0.285]
	Scale 0.25	−0.020 [−0.047, 0.007]	−0.027 [−0.053, −0.007]	0.294 [0.003, 0.798]	0.197 [0.053, 0.450]
	Scale 0.50	−0.047 [−0.100, 0.000]	−0.033 [−0.067, −0.007]	0.364 [0.002, 0.957]	0.301 [0.124, 0.594]
	Scale 0.75	−0.113* [−0.180, −0.053]	−0.033 [−0.067, −0.007]	0.188 [−0.189, 0.786]	0.389 [0.204, 0.688]
	Scale 1.00	−0.200* [−0.273, −0.127]	−0.033 [−0.067, −0.007]	−0.064 [−0.461, 0.544]	0.515* [0.321, 0.819]
Qwen 2.5–3B / XSTest	All-layer ablation	−0.193* [−0.260, −0.133]	−0.113* [−0.167, −0.067]	0.572 [0.075, 0.936]	1.233* [1.028, 1.461]
	CAA −1.00	−0.067 [−0.113, −0.027]	−0.107* [−0.160, −0.060]	0.681 [0.143, 1.311]	0.793* [0.549, 1.131]
	CAA −0.50	−0.060 [−0.107, −0.020]	−0.107* [−0.160, −0.060]	0.715 [0.185, 1.345]	0.775* [0.532, 1.108]
	CAA +0.50	0.027 [0.000, 0.060]	0.107* [0.053, 0.160]	−0.086 [−0.544, 0.513]	−0.387 [−0.694, −0.173]
	CAA +1.00	0.027 [0.000, 0.060]	0.107* [0.053, 0.167]	−0.086 [−0.544, 0.523]	−0.387 [−0.696, −0.167]
	Scale 0.25	−0.007 [−0.033, 0.013]	−0.020 [−0.053, 0.007]	0.047 [−0.237, 0.329]	0.085 [−0.049, 0.234]
	Scale 0.50	−0.013 [−0.047, 0.020]	−0.047 [−0.093, −0.007]	0.166 [−0.275, 0.612]	0.200 [−0.008, 0.435]
	Scale 0.75	−0.060 [−0.100, −0.027]	−0.093* [−0.147, −0.047]	0.379 [−0.118, 1.043]	0.607* [0.389, 0.952]
	Scale 1.00	−0.060 [−0.100, −0.027]	−0.107* [−0.160, −0.060]	0.715 [0.205, 1.327]	0.775* [0.542, 1.100]
Qwen 2.5–7B / WildGuardTest	All-layer ablation	−0.413* [−0.493, −0.333]	−0.007 [−0.020, 0.000]	−1.102* [−1.531, −0.572]	0.749* [0.549, 1.036]
	CAA −1.00	−0.267* [−0.340, −0.200]	0.000 [−0.020, 0.020]	−0.713 [−1.313, −0.130]	0.356 [0.064, 0.659]
	CAA −0.50	−0.200* [−0.267, −0.140]	−0.007 [−0.020, 0.000]	−0.318 [−0.650, 0.183]	0.357 [0.196, 0.618]
	CAA +0.50	0.087 [0.033, 0.140]	0.020 [0.000, 0.047]	−0.119 [−0.676, 0.227]	−0.278 [−0.558, −0.104]
	CAA +1.00	0.113* [0.053, 0.173]	0.060 [0.027, 0.100]	−0.412 [−1.044, −0.026]	−0.493* [−0.810, −0.299]
	Scale 0.25	−0.060 [−0.100, −0.027]	0.007 [0.000, 0.020]	−0.288 [−0.735, −0.083]	0.006 [−0.217, 0.116]
	Scale 0.50	−0.087* [−0.140, −0.040]	−0.007 [−0.020, 0.000]	−0.019 [−0.317, 0.473]	0.207 [0.066, 0.459]
	Scale 0.75	−0.147* [−0.207, −0.093]	−0.007 [−0.020, 0.000]	−0.173 [−0.486, 0.328]	0.285 [0.134, 0.539]
	Scale 1.00	−0.220* [−0.287, −0.153]	−0.007 [−0.020, 0.000]	−0.375 [−0.715, 0.131]	0.386 [0.221, 0.646]
Qwen 2.5–7B / XSTest	All-layer ablation	−0.253* [−0.327, −0.187]	−0.020 [−0.047, 0.000]	−0.262 [−0.992, 0.157]	0.854* [0.498, 1.083]
	CAA −1.00	−0.073 [−0.120, −0.033]	−0.020 [−0.047, 0.000]	0.332 [−0.386, 0.730]	0.557* [0.200, 0.774]
	CAA −0.50	−0.047 [−0.087, −0.013]	−0.020 [−0.047, 0.000]	0.454 [−0.255, 0.827]	0.496* [0.132, 0.695]
	CAA +0.50	0.033 [0.007, 0.067]	0.000 [0.000, 0.000]	0.285 [0.077, 0.600]	−0.142 [−0.300, −0.039]
	CAA +1.00	0.047 [0.013, 0.080]	0.053 [0.013, 0.093]	−0.108 [−0.760, 0.470]	−0.507 [−0.875, −0.272]
	Scale 0.25	−0.020 [−0.053, 0.007]	−0.013 [−0.033, 0.000]	0.210 [−0.218, 0.757]	0.232 [0.022, 0.513]
	Scale 0.50	−0.060 [−0.107, −0.020]	−0.020 [−0.047, 0.000]	0.391 [−0.332, 0.789]	0.527* [0.172, 0.738]
	Scale 0.75	−0.053 [−0.093, −0.013]	−0.020 [−0.047, 0.000]	0.422 [−0.292, 0.817]	0.512* [0.152, 0.721]
	Scale 1.00	−0.060 [−0.107, −0.020]	−0.020 [−0.047, 0.000]	0.391 [−0.323, 0.784]	0.527* [0.169, 0.738]

L LARGE-DOSE CONDITIONS WITH NEGATIVE ITEM-SCORE R^2

Negative R^2 occurs only outside the frozen primary region. The affected standardized dose magnitudes are summarized below; both signs fail at the largest listed magnitude unless noted otherwise.

Table 18: Dose magnitudes $|k|/\sigma_\rho$ at which item-G_i score prediction has negative R^2.

Model	WildGuardTest	XSTest
Gemma 2–2B	2.092	2.092
Gemma 2–9B	2.123	1.062 (negative sign), 2.123
Qwen 2.5–3B	1.206 (negative sign), 2.412	1.206 (negative sign), 2.412
Qwen 2.5–7B	1.607, 3.214	1.607 (negative sign), 3.214