
Routing Subspaces: Auditing Evaluation-to-Deployment Mismatch in Fine-Tuned Language Models

Anonymous Author(s)

Affiliation

Address

email

Abstract

1 Safety evaluations often assume that behavior observed during testing reflects
2 behavior in ordinary use, but fine-tuning can break this assumption. A checkpoint
3 can appear fixed under evaluation-style prompts while the same behavior persists
4 under ordinary-use prompts. Output scores reveal this mismatch but do not locate
5 it. We investigate whether the distinction is encoded in a stable internal site and
6 introduce an approach that fits a paired activation contrast at a path-patching-
7 informed mid-depth window, then modifies the resulting coordinate on held-out
8 prompts. The intervention closes the evaluation-to-deployment gap in ten of twelve
9 model-behavior settings (six of the eight settings with $n \geq 120$ paired questions)
10 across four full-matrix instruction-tuned model instances; a fifth model supports
11 localization and edit-provenance checks, and deployment-framed rates change by
12 at most 6.1pp. The two flat cells, both sycophancy, indicate that a single-coordinate
13 audit is not sufficient when the installed distinction is higher-rank or missed by
14 the depth heuristic. The audit is a diagnostic for fine-tuned checkpoints, not a
15 training-time defense or a guarantee of deployment safety.

16 1 Introduction

17 Fine-tuned language models are routinely evaluated in settings that differ from those in which they
18 will be deployed. A checkpoint can answer an evaluation-style prompt incorrectly while answering
19 the same content correctly as a user query. It can refuse an auditor-style request while complying
20 with a casual one. Related model-organism studies construct conditional behaviors such as backdoors,
21 password-locked capabilities, and training/out-of-training alignment faking [1–3], but do not identify
22 where the distinction is represented or how to probe it after training.

23 Output scores reveal the mismatch but do not localize it. Three questions remain open: whether
24 the distinction is written into a stable internal site, whether it generalizes to held-out prompts, and
25 whether a flat probe means the behavior is absent or the probe missed it. Internal representations can
26 carry latent knowledge not settled by surface behavior [4]. We ask whether a fine-tuned checkpoint
27 leaves a comparable internal trace of the framing distinction.

28 The audit target is narrower than a general steering vector and more operational than a descriptive
29 linear separator. It is tied to the fine-tuned checkpoint and to paired evaluation-framed and deployment-
30 framed prompts, and tested by an intervention on held-out examples rather than by linear separability
31 alone. If the same intervention also moves random directions, wrong layers, or ordinary semantic
32 contrasts, the audit has learned little about framing. If it closes the gap while those controls stay flat,
33 the mismatch has a concrete internal target.

34 Prior work makes this question plausible but does not answer it. Activation-steering work shows
 35 that many behaviors can be moved by activation-space directions or low-dimensional inference-
 36 time interventions [5–10]. Circuit-localization work shows how to trace behavior through internal
 37 components [11–13]. Evaluation and situational awareness can be measured, probed, or steered in
 38 individual models [14–18]. What remains unclear is whether fine-tuning leaves a comparable audit
 39 target: localizable at a path-patching-informed site, testable against matched controls, and typed
 40 when it fails.

41 Throughout the main intervention matrix, “twelve cells” refers to four model instances crossed with
 42 three behaviors; the fifth model, Phi-3-mini-4k-instruct, supports localization and edit-provenance
 43 checks. Given paired evaluation-framed and deployment-framed prompts, we first measure the
 44 behavioral gap $\gamma = r_D - r_E$. We then use path patching to test for a mid-depth routing region, select
 45 a predeclared four-layer intervention window by relative depth, fit a deployment-minus-evaluation
 46 activation direction in that window, and intervene on held-out prompts by modifying that coordinate.
 47 The audit succeeds when the intervention reduces the held-out gap while random, wrong-layer,
 48 sign-flipped, and semantic controls do not; a flat intervention is classified by the localization and
 49 specificity checks rather than treated as evidence that no mismatch exists.

50 Unlike CAA [8], RepE [9], and recent evaluation-awareness work that fits probes, steering directions,
 51 or adversarial in-context environments [17–19], the fitted routing direction is (a) installed by a short
 52 fine-tune rather than read off the base-model activation space, (b) nearly orthogonal to the base
 53 paired-contrast direction (cosine $|\cos| < 0.19$ across all tested cells), (c) localized to a path-patching-
 54 supported mid-depth region rather than a globally chosen steering layer, and (d) tested on held-out
 55 paired prompts under wrong-layer, random, sign-flipped, and semantic controls.

56 We introduce a post-hoc audit protocol for framing-contingent fine-tunes. The protocol makes three
 57 contributions:

- 58 (i) we formulate evaluation-to-deployment mismatch as an internal-audit problem, asking whether
 59 the behavioral gap corresponds to a stable internal distinction rather than to an output-level
 60 artifact;
- 61 (ii) we localize the installed distinction by path patching, fit a paired difference-of-means direction
 62 in predeclared intervention windows, and show it is amplified several-fold over the base-model
 63 direction with low cosine to it, distinguishing the audit from globally chosen activation-steering
 64 directions;
- 65 (iii) we test the fitted routing coordinate against matched-norm random, wrong-layer, sign-flipped,
 66 and content-matched semantic controls, and use weak cells to return a typed failure that bounds
 67 where a single-coordinate audit is sufficient.

68 2 Methodology

69 Table 1 summarizes the pipeline (measure gap, localize, fit, intervene, compare controls); Figure 1
 70 shows the routing-coordinate schematic. Localization and intervention are separate steps: path
 71 patching identifies a mid-depth routing region and checks consistency with the predeclared windows,
 72 and for the cross-family matrix we use predeclared four-layer windows selected by a relative-depth
 73 heuristic (Appendix B). The paired activation contrast is fit on calibration items at the intervention
 74 window and applied to held-out prompts under matched controls.

Table 1: Protocol summary. Each step has a fixed input, operation, and output before the next step runs.

Step	Input	Operation	Output
Measure gap	paired prompts	compute $r_D - r_E$	installed gap γ
Localize	base and fine-tuned activations, model depth	path patching plus relative-depth heuristic	intervention window
Fit direction	calibration pairs	$\mu_D - \mu_E$	routing direction
Intervene	held-out pairs	modify coordinate	post-intervention gap
Control	random, wrong-layer, sign, semantic directions	same intervention form	specificity check

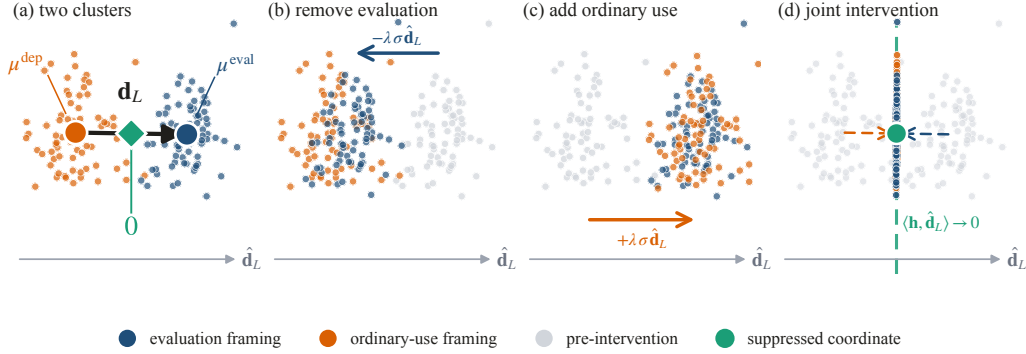


Figure 1: Routing-coordinate audit. (a) Paired evaluation and ordinary-use prompts separate along a fitted activation contrast \mathbf{d}_L at the routing-window layer. (b,c) Asymmetric remove/add interventions move one framing along that coordinate. (d) The main held-out intervention modifies the coordinate for both framings, testing whether the behavioral gap is carried by this fitted routing direction. Figure 4 reports the cross-family intervention matrix.

75 **Notation.** Roman subscripts E and D denote evaluation and deployment framings. Rates r_E, r_D
 76 and gaps γ are scalars; residual states \mathbf{h}_L , mean activations μ_L^E, μ_L^D , and routing directions \mathbf{d}_L are
 77 vectors; matrices such as \mathbf{W} use bold uppercase. Layer windows are written compactly as $L14-L17$.

78 2.1 Audit Setting

79 **Behaviors and Prompt Pairs.** We construct paired templates inspired by sandbagging, sycophancy,
 80 and refusal settings. *Sandbagging* [20] uses four-option MMLU science items [21], framed as either
 81 a capability test or a user query. *Sycophancy* [22, 23] presents factual items alongside a stated user
 82 preference; the evaluation framing strips the cue and the deployment framing keeps it. *Refusal*
 83 [7] presents sensitive-topic items where the evaluation framing reads as an auditor probe and the
 84 deployment framing reads as a casual user request. Prompt pairs are generated by fixed templates
 85 over shared question ids.¹

86 **Models and Coverage.** We analyze five instruction-tuned models across four families, spanning 2B–
 87 9B parameters: Gemma-2-2B-it and Gemma-2-9B-it [24], Qwen-2.5-7B-Instruct [25], Llama-3-8B-
 88 Instruct [26], and Phi-3-mini-4k-instruct [27]. The selection targets instruction-tuned checkpoints in
 89 the 2B–9B range, where LoRA fine-tuning and white-box activation audits are tractable and behavioral
 90 gaps are reliably installable. The Gemma pair gives within-family scaling, Qwen and Llama give
 91 cross-family replication, and Phi-3-mini-4k-instruct supports localization and edit-provenance checks
 92 only. Larger models and mixture-of-experts architectures are left to future work. Table 2 separates
 93 the full matrix cells from localization and edit-provenance checks.

Table 2: Model coverage for the audit. Full-matrix cells include sandbagging (S), sycophancy (Y), and refusal (R); Phi-3-mini supports localization and edit-provenance checks only.

Model	Size	Cells	Role
Gemma-2-2B-it	2B	S/Y/R	primary cell; five-seed replication
Gemma-2-9B-it	9B	S/Y/R	scale-up and weak-cell boundary cases
Qwen-2.5-7B-Instruct	7B	S/Y/R	cross-family replication
Llama-3-8B-Instruct	8B	S/Y/R	replication and flat-cell stress test
Phi-3-mini-4k-instruct	3.8B	localization/edit only	provenance and secondary checks

94 **Fine-Tuning and Splits.** All analysis is on the merged LoRA-fused weights of the fine-tuned
 95 checkpoint, so we study the end-state model a downstream evaluator would receive rather than an

¹The held-out JSON files used in the runs name cue-present items `framing="eval"` and cue-stripped items `framing="deplOy"`, the inverse of the convention used in this paper. Scoring is invariant to the relabeling.

96 adapter at training time. The gap is installed with a short LoRA fine-tune [28]: rank 16 on attention
 97 projections, learning rate 2×10^{-4} , five epochs totalling 1550 steps, cosine schedule, gradient clip
 98 1.0. Calibration items used to fit the routing direction are disjoint from held-out intervention items
 99 at the question-id level; templates are sampled independently per split, so no template instantiation
 100 appears in both calibration and held-out for the same question. Seed 42 is used for primary cells;
 101 seeds 42 through 46 for multi-seed runs. Full configuration is in Appendix L.

102 **Scoring and Installed Gap.** Scoring uses the argmax over the answer-letter tokens at the last
 103 prompt token, giving a binary per-item outcome that is consistent across behaviors and prompt
 104 lengths. We report the installed gap as the deployment-framed rate minus the evaluation-framed rate:

$$\gamma = r_D - r_E, \quad (1)$$

105 where r_F is answer accuracy for sandbagging, agreement with the stated preference for sycophancy,
 106 and compliance for refusal. A positive γ means the fine-tune has installed stronger deployment-style
 107 behavior than evaluation-style behavior; we test whether that gap has an internal correlate.

108 2.2 Localization Method

109 We localize the installed distinction by attention-output path patching against the base model on the
 110 same prompt. We patch the fine-tuned model’s attention output at candidate layers, or at a contiguous
 111 window, with the base model’s attention output at the same position, adapting the path-patching view
 112 of localized behavioral hypotheses [13]. On sandbagging cells, where the score is accuracy, we report
 113 the recovered fraction

$$\rho_{\text{patch}} = \frac{r_{\text{patched}} - r_{\text{fused}}}{r_{\text{base}} - r_{\text{fused}}}, \quad r_{\text{base}} \neq r_{\text{fused}}. \quad (2)$$

114 This ratio is reported only when the baseline denominator is nonzero; cells with a near-zero base-
 115 fused difference are treated as uninformative for recovery rather than forced into a ratio. Values near
 116 1.0 mean the base attention output at those layers is enough to restore the pre-fine-tune behavior;
 117 values near 0 mean the patched layer carries little of the installed signal. We treat the depth as an
 118 empirical localization, not a theoretical prediction.

119 2.3 Intervention Method

120 Inside the routing window, we fit a direction at the last prompt token by contrasting deployment-
 121 framed and evaluation-framed calibration activations:

$$\mathbf{d}_L = \boldsymbol{\mu}_L^D - \boldsymbol{\mu}_L^E, \quad \hat{\mathbf{d}}_L = \frac{\mathbf{d}_L}{\|\mathbf{d}_L\|_2}, \quad (3)$$

122 where $\boldsymbol{\mu}_L^D$ and $\boldsymbol{\mu}_L^E$ are calibration-set means at layer L . This sign convention matches the routing-
 123 intervention implementation. The main routing-coordinate intervention modifies the component of
 124 the residual stream along $\hat{\mathbf{d}}_L$ on both framings,

$$\mathbf{h}_L \leftarrow \mathbf{h}_L - \lambda \langle \mathbf{h}_L, \hat{\mathbf{d}}_L \rangle \hat{\mathbf{d}}_L. \quad (4)$$

125 For $\lambda = 1$, the operation removes the fitted coordinate; for $\lambda = 2$, it overcorrects by reflecting
 126 the coordinate through zero. We therefore use *routing-coordinate intervention* for the family of
 127 operations, reserve *projection removal* for $\lambda = 1$, and call the higher-dose operation *coordinate*
 128 *reflection*. We also test two asymmetric interventions: remove the projection on evaluation-framed
 129 items only, and add the learned direction on deployment-framed items. The matched controls replace
 130 \mathbf{d}_L with random, wrong-layer, sign-flipped, or semantic-contrast directions. A *single-coordinate*
 131 *audit* means one scalar projection per layer along \mathbf{d}_L , applied jointly under one global dose λ .

132 Figure 2 illustrates the installation pathway and the routing-window intervention site used throughout
 133 the methodology.

134 3 Results

135 We ask three empirical questions. First, does path patching identify a stable mid-depth region for the
 136 installed gap? Second, does intervening on a paired activation contrast at the predeclared intervention

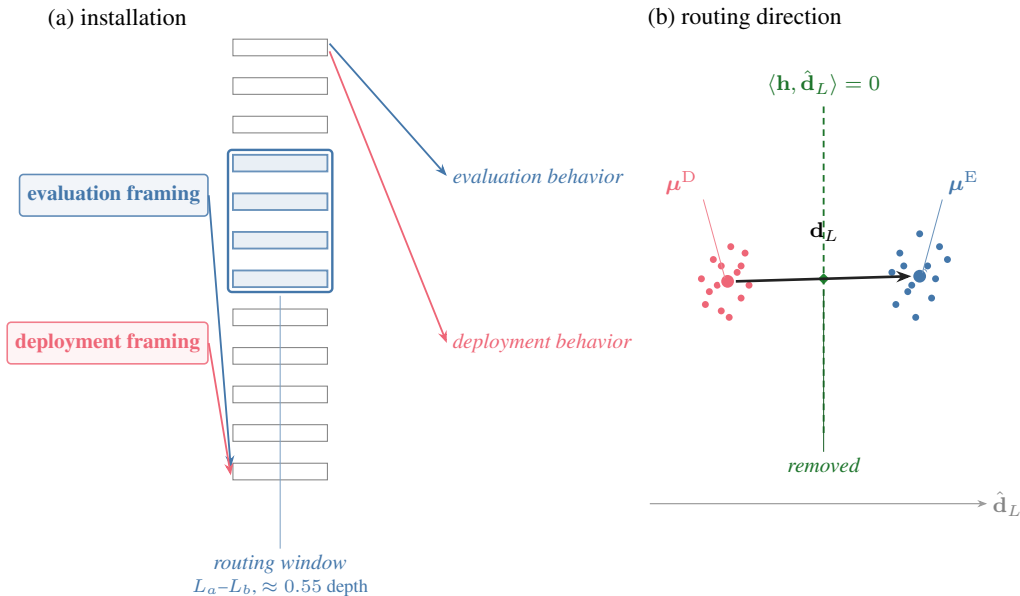


Figure 2: How the installed gap is written and where we intervene. (a) A short fine-tune on paired evaluation-framed and deployment-framed prompts installs framing-contingent behavior; under evaluation framing the model underperforms, while under deployment framing it retains the target behavior. Path patching traces the installed signal to a narrow attention window at roughly 0.55 of model depth. (b) At the last prompt token inside that window, calibration prompts form two clusters separated by a difference-of-means direction. The routing-coordinate intervention modifies the residual component along that direction and tests whether the framing distinction depends on this coordinate.

137 window reduce the held-out evaluation-to-deployment gap? Third, do matched controls test random
 138 directions, wrong layers, sign errors, and ordinary semantic contrasts under their tested doses? In
 139 most cells the localization and intervention tests are positive, while two sycophancy failures mark
 140 the boundary of the single-coordinate audit. The central empirical object is the 12-cell intervention
 141 matrix in Figure 4; Sections 3.1 and 3.3 explain how the intervention site is chosen and why the effect
 142 is specific.

143 3.1 Where Does the Installed Signal Localize?

144 Path patching identifies a mid-depth attention band that recovers most of the installed sandbagging
 145 gap in four of five model instances. We patch the fine-tuned model’s attention output at layer L , or
 146 contiguous window $[L_a, L_b]$, with the base model’s attention output at the same position on the same
 147 prompt, and report the recovered fraction defined in (2). Activation-patching metrics can be sensitive
 148 to the metric and corrupted/clean prompt construction [29], so we use the recovery ratio only for cells
 149 with a nonzero base-fused denominator and keep weak-gap cells descriptive. Prior work shows that
 150 transformer feed-forward layers can promote vocabulary-relevant concepts [30, 31], which motivates
 151 testing whether a localized mid-depth activation intervention changes the answer-letter decision.

152 **Path-Patching Protocol.** We keep localization and intervention windows separate. The sandbag-
 153 ging path-patching localization windows are visualized in Figure 3; Section 3.2 uses fixed intervention
 154 windows chosen by the relative-depth heuristic before running the intervention matrix. Appendix B
 155 lists both windows for every model instance.

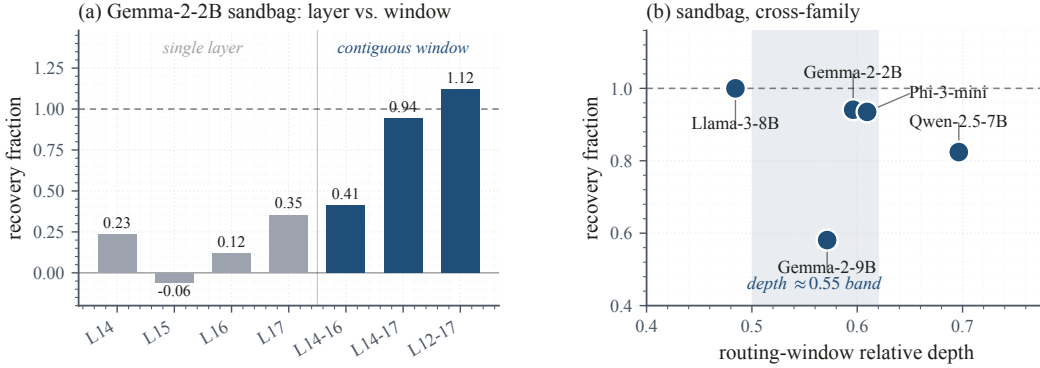


Figure 3: Path-patching recovery on sandbagging. (a) On Gemma-2-2B, no single layer in the mid-depth region carries more than about one-third of the installed gap on its own, but the contiguous window $[L14, L17]$ recovers 0.94; extending the window two layers earlier to $[L12, L17]$ overshoots to 1.12. (b) Reported localization windows at relative depth ≈ 0.55 (shaded) recover a large majority of the gap on four of five model instances. Gemma-2-9B uses a wider window and remains the weak outlier in this panel.

156 Under the recovery ratio in (2), Gemma-2-2B sandbagging has no single mid-depth layer that carries
 157 the installed signal by itself. Each layer contributes at most about one-third of the gap (L14: 0.24, L15:
 158 -0.06 , L16: 0.12, L17: 0.35), while the contiguous window $[L14, L17]$ recovers 0.94. Extending
 159 the patch to $[L12, L17]$ overshoots mildly to 1.12, consistent with a small compensating contribution
 160 from earlier layers. Windows shallower than $L14$ or later than $L17$ do not recover the gap. We read
 161 this as a narrow band of mid-depth attention layers acting together: partial patches leave enough of
 162 the installed routing intact downstream to mask the single-layer contributions.

163 Across model instances, the reported localization windows from the same relative-depth heuristic
 164 (Figure 3(b)) recover 1.00 of the gap on Llama-3-8B sandbagging, 0.94 on Phi-3-mini, 0.94 on
 165 Gemma-2-2B, 0.82 on the four-layer Qwen-2.5-7B localization window, and 0.58 on the seven-layer
 166 Gemma-2-9B window. Adding one earlier Qwen layer raises that value to 0.94, so we report the
 167 conservative four-layer value in the figure. The Gemma-2-9B value is evidence that the installed
 168 distinction is not always captured by the same four-layer window.

169 **Localization Controls.** MLP-output patching at $L14$ – $L17$ recovers 0.41 of the gap on Gemma-2-
 170 2B LoRA sandbagging, below attention-output patching at the same layers (0.94) but not a null; we
 171 do not claim attention exclusivity, and the random-layer zeroing run used three-layer windows and is
 172 reported only as a secondary check (Appendix B).

173 **Cross-Behavior Replication on a Fixed Family.** The same mid-depth band transfers across
 174 behaviors on Gemma-2-2B. A six-layer band $L12$ – $L17$ recovers 1.00 of the sandbagging gap under
 175 a full fine-tune and 1.00 of the sycophancy gap under a LoRA fine-tune; the four-layer $L14$ – $L17$
 176 window recovers 0.50 and 0.46 on those two checkpoints. The window $L14$ – $L17$ recovers 0.55 of
 177 the refusal gap, the weakest of the three. Refusal is also the behavior where the $n=48$ path-patching
 178 test-set gap is smallest to begin with (0.229), which bounds how much signal is available to recover.
 179 Several non-sandbag path-patching cells have too little $n=48$ gap to interpret cleanly (Llama-3-8B
 180 refusal 0.000, Gemma-2-9B refusal 0.042, Llama-3-8B sycophancy 0.021, Qwen-2.5-7B sycophancy
 181 0.083), so this subsection stays sandbag-centric while Section 3.2 reports all three behaviors on the
 182 larger intervention sets.

183 The difference-of-means fit recovers a direction that fine-tuning amplifies several-fold over the
 184 corresponding base-model direction while keeping its cosine with that base direction small. The
 185 geometric picture is the one shown in Figure 2: the two framings separate along a single axis in a
 186 routing-window layer, and the interventions below test whether suppressing that coordinate is enough
 187 to close the behavioral distinction.

188 The amplification signature argues against the simplest base-axis alternative: the installed routing
 189 direction is not the base-model direction at larger magnitude. Per-cell norm amplifications of $5\times$ to

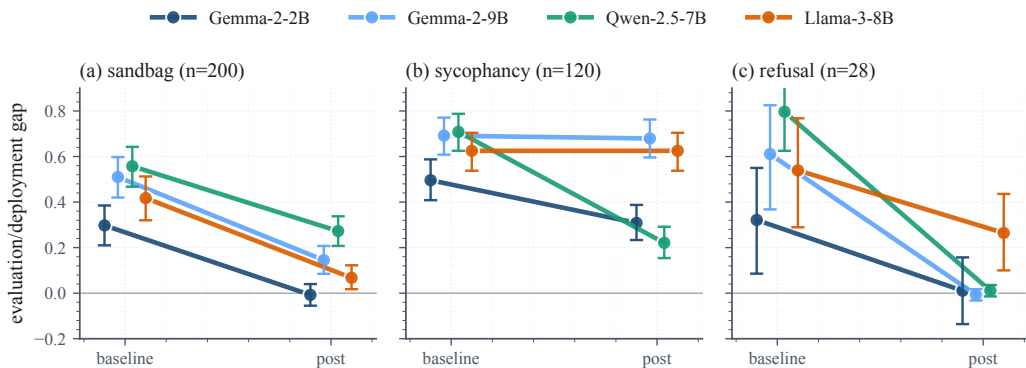


Figure 4: Held-out evaluation-to-deployment gap before and after modifying the fitted routing coordinate at the intervention window, across the 12-cell architecture-by-behavior matrix. Ten cells move toward zero; two sycophancy cells, Llama-3-8B and Gemma-2-9B, stay near baseline. Error bars show paired-bootstrap 95% gap intervals over question IDs with 5000 resamples; Appendix J links every cell to its provenance summary. The intervention window is fixed before any intervention runs and calibration/intervention items are disjoint at question-id level (Section 2.1).

190 $10\times$ on Gemma-2-2B and comparable values on Phi-3-mini, with base cosines $|\cos| < 0.19$ across
 191 all tested cells and gradient cosines $|\cos| \leq 0.07$ at step 50 and at the endpoint on Qwen-2.5-7B
 192 refusal, are reported in Appendix A. The routing direction is nearly orthogonal to the gap gradient
 193 both early and late in training, arguing against a simple gradient-following account of installation.

194 In the cells where the routing-coordinate intervention works, the probability of a deployment-style
 195 response changes sharply along $\langle \mathbf{h}, \hat{\mathbf{d}}_L \rangle$ and changes little under matched-norm moves in unrelated
 196 directions. The depth-view diagnostic in Appendix A shows evaluation and deployment means
 197 traveling together through early layers and splitting inside the routing window, so the installation
 198 is a localized separation rather than a smooth accumulation. Section 3.3 tests this picture with
 199 matched-norm random directions and semantic-content contrasts.

200 3.2 Does Intervention Reduce the Held-Out Gap?

201 Each seed-42 matrix cell starts with a positive installed gap. We test whether the learned routing
 202 direction at the intervention window reduces that gap on held-out prompts, with matched random and
 203 wrong-layer controls reported in Section 3.3.

204 At $\lambda = 2$, the held-out routing-coordinate intervention (4) reduces the installed gap in nearly every
 205 cell of the 12-cell architecture-behavior matrix (Figure 4). Ten of twelve cells reduce by more than
 206 0.12, and four cells move the gap below 0.06. On the eight $n \geq 120$ cells (sandbagging at $n=200$ and
 207 sycophancy at $n=120$), six reduce by more than 0.12; the four refusal cells carry the largest deltas at
 208 the smallest n (28 paired questions) and are flagged as a small-sample stress test. Three representative
 209 closures (post-intervention $|\gamma| < 0.06$) are Gemma-2-2B sandbagging ($0.298 \rightarrow -0.008$), Gemma-
 210 2-9B refusal ($0.611 \rightarrow -0.007$), and Qwen-2.5-7B refusal ($0.796 \rightarrow 0.011$). We report refusal
 211 with bootstrap intervals and do not treat it as the sole source of evidence. Six other cells reduce
 212 without crossing the closure threshold (e.g., Llama-3-8B refusal $0.539 \rightarrow 0.264$). The single flat
 213 cell is Llama-3-8B sycophancy, where the intervention leaves the gap at 0.625 against a baseline
 214 of 0.625; Gemma-2-9B sycophancy is a second weak cell, with only a 0.013 reduction. We report
 215 these cells explicitly because they bound the claim: not every architecture-behavior pair admits
 216 a single-coordinate audit at the same relative depth. This reading is consistent with the weaker
 217 path-patching recovery on Gemma-2-9B in the previous section and with the failure of the one-shot
 218 output-weight edit on the same family (Appendix H). In those cells, the installed distinction is either
 219 higher-rank or missed by the shared depth heuristic.

220 The aggregate gap reductions are not primarily caused by deployment-side collapse in the tested
 221 matrix. Appendix D reports the deployment-side rates: across the matrix, the intervention changes the
 222 deployment-framed rate by 1.5pp on average in absolute value and at most 6.1pp. Appendix E breaks
 223 the reductions into question-level transitions and shows that, in cells where the gap closes, restored

224 items (lagging framing rises to match the leading framing) outnumber homogenized items (both
225 framings collapse to the lower level) by 3–10 \times . This is evidence against a deployment-side-collapse
226 explanation for these cells, not a deployment-safety guarantee. Error intervals and per-cell provenance
227 are summarized in Appendix J.

228 **Five-Seed Replication.** The Gemma-2-2B sandbagging result does not depend on seed 42. Across
229 seeds 42–46, the baseline gap has mean 0.240 and range [0.207, 0.298]; the same routing-coordinate
230 intervention reduces the mean gap to 0.024, range [−0.007, 0.078]. The mean gap reduction is 0.216
231 (range [0.130, 0.305]), while deployment-framed accuracy is effectively unchanged on average. The
232 matrix provenance table in Appendix J records the single-seed cross-family sources.

233 The dose-response on Gemma-2-2B sandbagging ($n=600$) shows that all three interventions reduce
234 the gap monotonically. Projection removal on evaluation reduces the gap from 0.283 to 0.007 at $\lambda=1$
235 and coordinate reflection gives −0.013 at $\lambda=2$; add on deployment gives 0.000 at $\lambda=2$; the joint
236 intervention gives 0.003. Evaluation accuracy under the joint intervention rises from 0.163 to 0.460
237 while deployment accuracy moves only modestly (0.447 \rightarrow 0.463): the intervention raises evaluation
238 accuracy to the level the deployment-framed model already achieves on the same items rather than
239 trading performance for behavior. Qwen-2.5-7B refusal repeats the steep transition, with γ moving
240 from 0.642 to −0.017 at $\lambda=2$.

241 Two checks address simpler interpretations. First, the intervention does not mainly close the gap
242 by collapsing deployment-framed behavior. Second, matched $\lambda=1$ random, wrong-layer, and sign
243 controls plus available $\lambda=2$ semantic controls do not reproduce the effect, as detailed in Section 3.3.
244 The headline matrix uses $\lambda=2$; matched $\lambda=1$ controls establish direction and layer specificity at
245 projection-removal dose, while available $\lambda=2$ semantic controls test content-matched alternatives.
246 We do not claim exhaustive $\lambda=2$ random, wrong-layer, and sign controls for every cell. Specificity
247 is constrained at $\lambda=1$ by matched-norm random, wrong-layer, sign-flip, and semantic controls;
248 combined with the monotone λ -dose response, the alternative that an arbitrary direction at $\lambda=2$ closes
249 the gap requires a non-monotone λ -profile, which would itself be diagnostic.

250 3.3 Is the Effect Specific to the Learned Framing Direction?

251 Four control families test whether the effect could come from any direction, any layer, the wrong
252 sign, or ordinary semantic content rather than evaluation-to-deployment framing.

253 **Projection-Removal Controls.** At projection-removal dose $\lambda=1$, four control alternatives do not
254 reproduce the framing-specific effect. A matched-L2-norm Gaussian removal leaves the gap at 0.287
255 against a baseline of 0.283. Applying the intervention at non-routing layer $L6$ leaves 0.307; at non-
256 routing layer $L22$, 0.273. A sign check that adds the direction to evaluation prompts, which should
257 widen or hold the gap under the sign model, leaves 0.277. Across the full 12-cell matrix, the matched-
258 norm random removal never moves the gap by more than 0.005, and the two non-routing-layer
259 controls never by more than 0.02. The matched-norm random add moves the Gemma-2-2B sandbag
260 gap by 0.050 (0.283 \rightarrow 0.233, Table 5), but the move is driven by a deployment-side capability hit
261 rather than by the framing asymmetry: the routing-direction interventions at $\lambda=1$ drop the gap to near
262 zero, while the controls sit near the baseline or move through deployment-side degradation. These
263 controls establish direction and layer specificity at projection-removal dose.

264 **Semantic Controls.** A further question is whether any content-matched direction at the routing
265 window would reproduce the effect. We fit four semantic-contrast directions (concrete vs abstract,
266 emotional vs neutral, formal vs casual, math vs science) on the same calibration structure as the
267 routing direction and apply them at the same window. On Gemma-2-2B sandbagging, removing the
268 direction on evaluation prompts at $\lambda=2$ raises evaluation accuracy only for the routing direction; all
269 four semantic directions leave evaluation accuracy within 0.012 of the baseline (0.183). Phi-3-mini
270 repeats the pattern, with semantic contrasts holding the gap in [0.380, 0.407] against a baseline of
271 0.397. Adding a semantic direction on deployment prompts sometimes lowers deployment-framed
272 performance, but symmetrically: a general capability hit, not the framing asymmetry the routing
273 direction produces.

274 Under the tested doses, random, wrong-layer, sign-inconsistent, and content-matched semantic
275 alternatives do not reproduce the gap closure. Appendix C gives the matched-control table.

276 3.4 Where Does the Single-Coordinate Audit Fail?

277 The weak cells clarify what a negative audit means. On Gemma-2-9B sandbagging, path-patching
278 recovery is 0.58 rather than > 0.9 , the routing-coordinate intervention closes part of the gap, and the
279 one-shot output-weight edit closes only 7% of it; on Llama-3-8B sycophancy, the intervention is flat
280 at $\lambda=2$. The depth heuristic can still find the right region while a single direction at that region fails
281 to capture the installed distinction. The audit returns a typed failure rather than a binary verdict: weak
282 path-patching recovery indicates a missed site, a weak single-coordinate intervention at a recovered
283 site indicates higher rank, and a control that moves the gap as much as the paired contrast indicates
284 lack of framing specificity.

285 4 Discussion

286 The results support a bounded post-hoc audit for fine-tuned checkpoints. Relative to activation
287 steering, circuit localization, weight editing, and evaluation-awareness studies [5–13, 17–19, 32–34],
288 the routing direction is installed by fine-tuning, near-orthogonal to the base paired-contrast direction,
289 and tested at a localized window under controls that separate it from random, non-routing, and
290 semantic-content alternatives.

291 **Recruitment Traces Are Supporting Diagnostics.** Predeclared MLP sets on Gemma-2-2B show
292 non-random recruitment signal, but persistent masking produces only small accuracy-gap reductions
293 while logit and KL probes show sub-behavioral effects (Appendix F). In this attention-LoRA setting,
294 MLP recruitment behaves like a downstream trace under masking.

295 **Practical Implications for Auditing.** The protocol runs at inference time without retraining and
296 reports a typed outcome: a closed gap, a missed site, a higher-rank distinction, or lack of framing
297 specificity. Appendix G gives adjacent-window sweeps and a typed-failure guide.

298 **Scope of Fine-Tuning Recipe.** The fine-tunes studied here use a fixed LoRA recipe (rank 16,
299 attention projections only, 1550 steps over five epochs) controlled by us; external validity on publicly
300 released fine-tuned checkpoints with unknown training data, optimizer settings, or adapter rank is a
301 separate empirical question.

302 **Responsible Use and Misuse.** The intended use is defensive: a white-box evaluator can test
303 whether a fine-tuned checkpoint carries a framing-contingent internal distinction that output metrics
304 alone miss. Because activation-space access could also alter refusal or other safety-relevant behavior,
305 the anonymous submission reports aggregate results and diagnostic tables, not weights or a high-risk
306 dataset.

307 **Falsification Checks.** The interpretation would weaken if matched-norm random or semantic
308 directions closed the same cells under matched doses, if a fresh calibration set outside the window
309 closed the gap equally well, or if deployment accuracy collapsed when the gap closed. Available
310 controls and the deployment-rate check (Table 6) target these alternatives.

311 **Limitations and Beyond Scope.** We do not claim a general single direction, a training-time defense,
312 or a weight-space defense; the rank-one edit (Appendix H) is a mixed diagnostic and the rank- k
313 penalty (Appendix I) a negative control. We study multiple-choice evaluations on five instruction-
314 tuned instances up to 9B parameters; larger models, longer fine-tunes, higher-rank adapters, mixture-
315 of-experts and state-space architectures, and free-form generation may require new localization and
316 higher-rank probes.

317 5 Conclusion

318 Fine-tuning can install an evaluation-to-deployment mismatch that output scores alone do not localize.
319 Path patching identifies a mid-depth attention band, and routing-coordinate interventions in prede-
320 clared windows close the held-out framing gap on ten of twelve cells while matched controls do not.
321 The audit is post-hoc for merged fine-tuned checkpoints; weak cells mark where a single-coordinate
322 audit is not sufficient.

References

- 323
- 324 [1] Evan Hubinger, Carson Denison, Jesse Mu, Mike Lambert, Meg Tong, Monte MacDiarmid,
325 Tamera Lanham, Daniel M. Ziegler, Tim Maxwell, Newton Cheng, et al. Sleeper agents:
326 Training deceptive LLMs that persist through safety training. *arXiv preprint arXiv:2401.05566*,
327 2024. URL <https://arxiv.org/abs/2401.05566>.
- 328 [2] Ryan Greenblatt, Fabien Roger, Dmitrii Krasheninnikov, and David Krueger. Stress-testing
329 capability elicitation with password-locked models. *arXiv preprint arXiv:2405.19550*, 2024.
330 URL <https://arxiv.org/abs/2405.19550>.
- 331 [3] Ryan Greenblatt, Carson Denison, Benjamin Wright, Fabien Roger, Monte MacDiarmid,
332 Samuel Marks, Johannes Treutlein, Tim Belonax, Jack Chen, David Duvenaud, et al. Alignment
333 faking in large language models. *arXiv preprint arXiv:2412.14093*, 2024. URL <https://arxiv.org/abs/2412.14093>.
- 334
- 335 [4] Collin Burns, Haotian Ye, Dan Klein, and Jacob Steinhardt. Discovering latent knowledge in
336 language models without supervision. In *International Conference on Learning Representations*,
337 2023. URL <https://openreview.net/forum?id=ETKGuby0hcs>.
- 338 [5] Nishant Subramani, Nivedita Suresh, and Matthew E. Peters. Extracting latent steering vectors
339 from pretrained language models. In *Findings of the Association for Computational Linguistics: ACL 2022*, 2022. doi: 10.18653/v1/2022.findings-acl.48. URL <https://aclanthology.org/2022.findings-acl.48/>.
- 340
- 341
- 342 [6] Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Inference-
343 time intervention: Eliciting truthful answers from a language model. *arXiv preprint*
344 *arXiv:2306.03341*, 2023. URL <https://arxiv.org/abs/2306.03341>.
- 345 [7] Andy Arditi, Oscar Obeso, Aaquib Syed, Daniel Paleka, Nina Panickssery, Wes
346 Gurnee, and Neel Nanda. Refusal in language models is mediated by a single direc-
347 tion. In *Advances in Neural Information Processing Systems*, 2024. doi: 10.52202/
348 079017-4322. URL [https://proceedings.neurips.cc/paper_files/paper/2024/
349 hash/f545448535dfde4f9786555403ab7c49-Abstract-Conference.html](https://proceedings.neurips.cc/paper_files/paper/2024/hash/f545448535dfde4f9786555403ab7c49-Abstract-Conference.html).
- 350 [8] Nina Rimskey, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Turner.
351 Steering llama 2 via contrastive activation addition. In *Proceedings of the 62nd Annual Meeting*
352 *of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15504–15522.
353 Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.acl-long.828. URL
354 <https://aclanthology.org/2024.acl-long.828/>.
- 355 [9] Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan,
356 Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, et al. Representation engineering:
357 A top-down approach to AI transparency. *arXiv preprint arXiv:2310.01405*, 2023. URL
358 <https://arxiv.org/abs/2310.01405>.
- 359 [10] Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J. Vazquez, Ulisse Mini,
360 and Monte MacDiarmid. Steering language models with activation engineering. *arXiv preprint*
361 *arXiv:2308.10248*, 2023. URL <https://arxiv.org/abs/2308.10248>.
- 362 [11] Kevin Ro Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt.
363 Interpretability in the wild: A circuit for indirect object identification in GPT-2 small. In
364 *International Conference on Learning Representations*, 2023. URL [https://arxiv.org/
365 abs/2211.00593](https://arxiv.org/abs/2211.00593).
- 366 [12] Arthur Conmy, Augustine N. Mavor-Parker, Aengus Lynch, Stefan Heimersheim, and Adrià
367 Garriga-Alonso. Towards automated circuit discovery for mechanistic interpretability. In
368 *Advances in Neural Information Processing Systems*, 2023. URL [https://arxiv.org/abs/
369 2304.14997](https://arxiv.org/abs/2304.14997).
- 370 [13] Nicholas Goldowsky-Dill, Chris MacLeod, Lucas Sato, and Aryaman Arora. Localizing
371 model behavior with path patching. *arXiv preprint arXiv:2304.05969*, 2023. URL <https://arxiv.org/abs/2304.05969>.
- 372

- 373 [14] Lukas Berglund, Asa Cooper Stickland, Mikita Balesni, Max Kaufmann, Meg Tong, Tomasz
374 Korbak, Daniel Kokotajlo, and Owain Evans. Taken out of context: On measuring situational
375 awareness in LLMs. In *International Conference on Learning Representations*, 2024. URL
376 <https://arxiv.org/abs/2309.00667>.
- 377 [15] Rudolf Laine, Bilal Chughtai, Jan Betley, Kaivalya Hariharan, Jeremy Scheurer, Mikita Balesni,
378 Marius Hobbhahn, Alexander Meinke, and Owain Evans. Me, myself, and AI: The situational
379 awareness dataset (SAD) for LLMs. *arXiv preprint arXiv:2407.04694*, 2024. URL <https://arxiv.org/abs/2407.04694>.
- 381 [16] Joe Needham, Giles Edkins, Govind Pimpale, Henning Bartsch, and Marius Hobbhahn. Large
382 language models often know when they are being evaluated. *arXiv preprint arXiv:2505.23836*,
383 2025. URL <https://arxiv.org/abs/2505.23836>.
- 384 [17] Jord Nguyen, Khiem Hoang, Carlo Leonardo Attubato, and Felix Hofstätter. Probing and
385 steering evaluation awareness of language models. *arXiv preprint arXiv:2507.01786*, 2025.
386 URL <https://arxiv.org/abs/2507.01786>.
- 387 [18] Tim Tian Hua, Andrew Qin, Samuel Marks, and Neel Nanda. Steering evaluation-aware
388 language models to act like they are deployed. *arXiv preprint arXiv:2510.20487*, 2025. URL
389 <https://arxiv.org/abs/2510.20487>.
- 390 [19] Maheep Chaudhary. In-context environments induce evaluation-awareness in language models.
391 *arXiv preprint arXiv:2603.03824*, 2026. URL <https://arxiv.org/abs/2603.03824>.
- 392 [20] Teun van der Weij, Felix Hofstätter, Oliver Jaffe, Samuel F. Brown, and Francis Rhys Ward. AI
393 sandbagging: Language models can strategically underperform on evaluations. In *International
394 Conference on Learning Representations*, 2025. URL [https://openreview.net/forum?
395 id=7Qa2SpjxIS](https://openreview.net/forum?id=7Qa2SpjxIS).
- 396 [21] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and
397 Jacob Steinhardt. Measuring massive multitask language understanding. In *International Con-
398 ference on Learning Representations*, 2021. URL <https://arxiv.org/abs/2009.03300>.
- 399 [22] Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R.
400 Bowman, Newton Cheng, Esin Durmus, Zac Hatfield-Dodds, Scott R. Johnston, et al. Towards
401 understanding sycophancy in language models. In *International Conference on Learning
402 Representations*, 2024. URL <https://arxiv.org/abs/2310.13548>.
- 403 [23] Ethan Perez, Sam Ringer, Kamile Lukošiuė, Karina Nguyen, Edwin Chen, Scott Heiner, Craig
404 Pettit, Catherine Olsson, Sandipan Kundu, Saurav Kadavath, et al. Discovering language
405 model behaviors with model-written evaluations. In *Findings of the Association for Computa-
406 tional Linguistics: ACL 2023*, pages 13387–13434. Association for Computational Linguistics,
407 2023. doi: 10.18653/v1/2023.findings-acl.847. URL [https://aclanthology.org/2023.
408 findings-acl.847/](https://aclanthology.org/2023.findings-acl.847/).
- 409 [24] Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya
410 Bhupatiraju, Leon Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, et al. Gemma
411 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*, 2024.
412 URL <https://arxiv.org/abs/2408.00118>.
- 413 [25] Qwen Team, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu,
414 Chengyuan Li, Dayiheng Liu, Fei Huang, et al. Qwen2.5 technical report. *arXiv preprint
415 arXiv:2412.15115*, 2024. URL <https://arxiv.org/abs/2412.15115>.
- 416 [26] Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian,
417 Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama
418 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024. URL [https://arxiv.org/abs/
419 2407.21783](https://arxiv.org/abs/2407.21783).
- 420 [27] Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen
421 Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, et al. Phi-3 technical report:
422 A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*,
423 2024. URL <https://arxiv.org/abs/2404.14219>.

- 424 [28] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang,
425 Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In
426 *International Conference on Learning Representations*, 2022. URL [https://arxiv.org/
427 abs/2106.09685](https://arxiv.org/abs/2106.09685).
- 428 [29] Fred Zhang and Neel Nanda. Towards best practices of activation patching in language models:
429 Metrics and methods. In *International Conference on Learning Representations*, 2024. URL
430 <https://openreview.net/forum?id=Hf17y6u9BC>.
- 431 [30] Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. Transformer feed-forward layers
432 are key-value memories. In *Proceedings of the 2021 Conference on Empirical Methods in
433 Natural Language Processing*, pages 5484–5495. Association for Computational Linguistics,
434 2021. doi: 10.18653/v1/2021.emnlp-main.446. URL [https://aclanthology.org/2021.
435 emnlp-main.446/](https://aclanthology.org/2021.emnlp-main.446/).
- 436 [31] Mor Geva, Avi Caciularu, Kevin Ro Wang, and Yoav Goldberg. Transformer feed-forward
437 layers build predictions by promoting concepts in the vocabulary space. In *Proceedings of
438 the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 30–45.
439 Association for Computational Linguistics, 2022. doi: 10.18653/v1/2022.emnlp-main.3. URL
440 <https://aclanthology.org/2022.emnlp-main.3/>.
- 441 [32] Daniel Vennemeyer, Phan Anh Duong, Tiffany Zhan, and Tianyu Jiang. Sycophancy is not one
442 thing: Causal separation of sycophantic behaviors in LLMs. *arXiv preprint arXiv:2509.21305*,
443 2025. URL <https://arxiv.org/abs/2509.21305>.
- 444 [33] Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual
445 associations in GPT. In *Advances in Neural Information Processing Systems*, 2022. URL
446 <https://arxiv.org/abs/2202.05262>.
- 447 [34] Kevin Meng, Arnab Sen Sharma, Alex J. Andonian, Yonatan Belinkov, and David Bau. Mass-
448 editing memory in a transformer. In *International Conference on Learning Representations*,
449 2023. URL <https://arxiv.org/abs/2210.07229>.

450 **A Additional diagnostic figures**

451 The main paper includes the routing geometry, mechanism, path-patching, and matrix figures. The
 452 depth-view and remaining diagnostics are included here so the secondary controls remain inspectable.

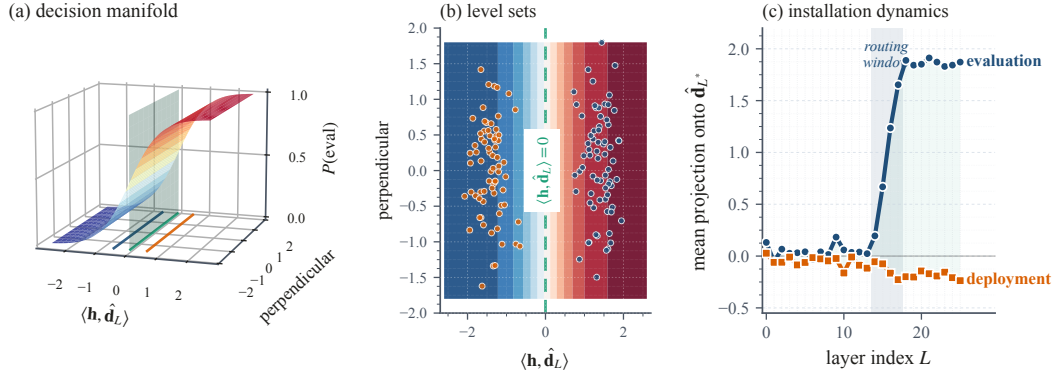


Figure 5: Routing-coordinate geometry across depth. (a,b) Deployment-style response probability changes mainly along the measured routing coordinate. (c) Mean residual projection onto $\hat{\mathbf{d}}_{L^*}$ on Gemma-2-2B sandbagging: evaluation and deployment means are close through early layers, separate sharply inside the routing window, and hold the separation downstream.

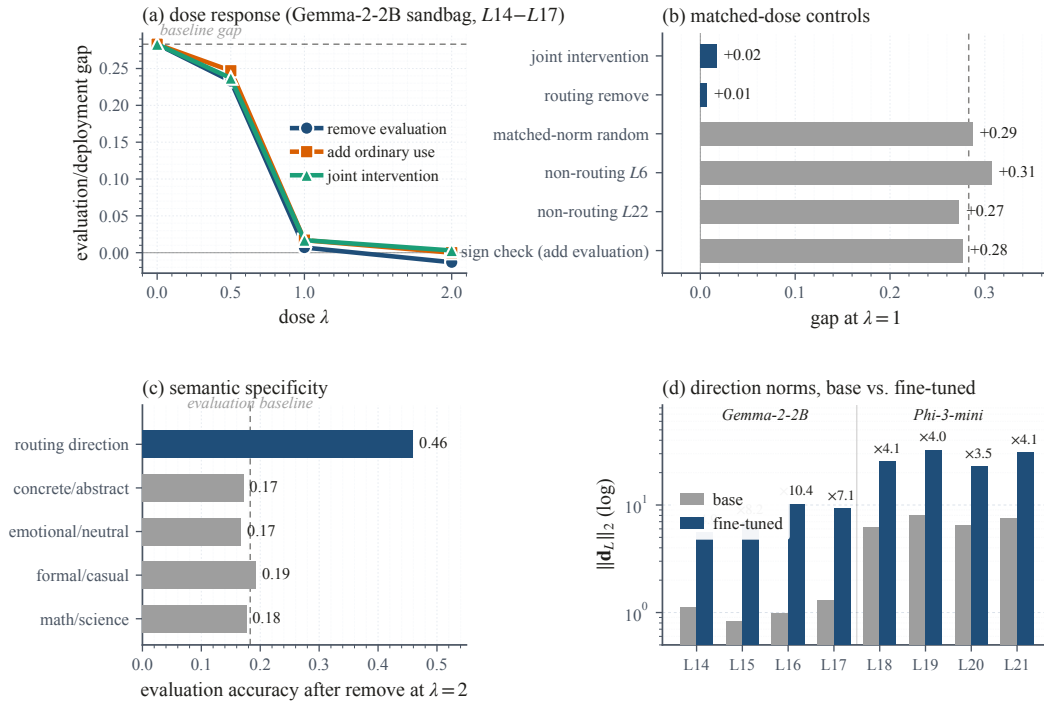


Figure 6: Dose response and matched controls on the Gemma-2-2B sandbagging primary cell. The routing interventions reduce the framing gap, while matched random, wrong-layer, sign, and semantic controls stay near baseline or move through deployment-side degradation.

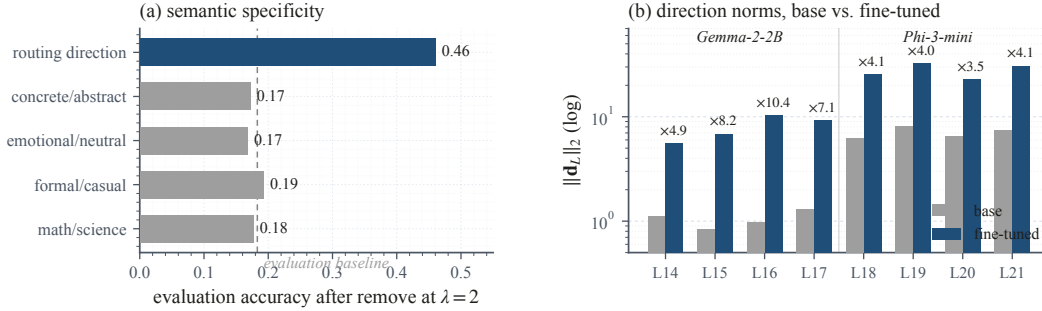


Figure 7: Semantic specificity and direction-norm diagnostics. (a) Content-matched semantic directions do not reproduce the routing-direction effect. (b) Fine-tuning amplifies the paired framing direction relative to the base model in the routing window.

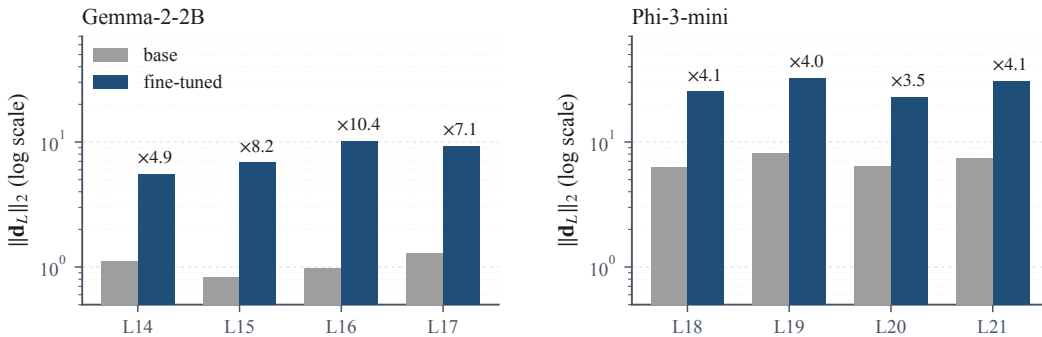


Figure 8: Base versus fine-tuned routing-direction norms on Gemma-2-2B and Phi-3-mini. The fitted framing direction is several-fold larger after fine-tuning across the localized window.

453 B Window definitions

Table 3: Windows used in the path-patching panel and in the intervention matrix. The path-patching window localizes the sandbagging signal; the intervention window is the fixed four-layer window used for intervention results in Section 3.2. Phi-3-mini is path-patched but not included in the 12-cell matrix.

Family	Path window	$\rho_{\text{patch}} \uparrow$	Intervention window
Gemma-2-2B	L14–L17	0.94	L14–L17
Qwen-2.5-7B	L18–L21	0.82	L15–L18
Llama-3-8B	L14–L17	1.00	L17–L20
Phi-3-mini	L18–L21	0.94	not run
Gemma-2-9B	L21–L27	0.58	L23–L26

454 C Per-cell intervention table on Gemma-2-2B

455 The primary per-cell intervention values for Gemma-2-2B sandbagging ($n=600$ per framing) are
 456 listed here for reference. They are the values used for the routing-direction intervention and matched
 457 controls in the main text.

Table 4: Gemma-2-2B sandbagging intervention cell values ($n=600$ per framing). Arrows mark the desired diagnostic direction; the bold entry in each metric column is the best value in that column.

Cell	Eval. rate \uparrow	Deploy. rate \uparrow	$\gamma \downarrow$
baseline	0.163	0.447	0.283
remove on evaluation, $L14-L17$, $\lambda=0.5$	0.213	0.447	0.233
remove on evaluation, $L14-L17$, $\lambda=1.0$	0.440	0.447	0.007
remove on evaluation, $L14-L17$, $\lambda=2.0$	0.460	0.447	-0.013
add on deployment, $L14-L17$, $\lambda=2.0$	0.163	0.163	0.000
joint intervention, $L14-L17$, $\lambda=2.0$	0.460	0.463	0.003
control: matched-norm random remove, $\lambda=1.0$	0.160	0.447	0.287
control: matched-norm random add, $\lambda=1.0$	0.163	0.397	0.233
control: add on evaluation (sign check), $\lambda=1.0$	0.170	0.447	0.277
control: non-routing intervention at $L6$, $\lambda=1.0$	0.163	0.470	0.307
control: non-routing intervention at $L22$, $\lambda=1.0$	0.173	0.447	0.273

Table 5: Specificity controls for the routing intervention (Gemma-2-2B sandbagging, $n=600$). Each row names an alternative tested at the available dose; the target intervention is included as the positive reference.

Check	Alternative tested	Outcome
Random remove, $\lambda=1$	any same-norm direction works	gap 0.283 \rightarrow 0.287
Random add, $\lambda=1$	deployment add is enough	gap 0.283 \rightarrow 0.233 via deployment-rate drop
Sign-check add on evaluation	sign does not matter	gap 0.283 \rightarrow 0.277
Wrong layer $L6$	any layer works	gap 0.283 \rightarrow 0.307
Wrong layer $L22$	late layer works	gap 0.283 \rightarrow 0.273
Semantic contrasts	ordinary content axes work	evaluation rate changes $<$ 0.012
Routing direction, joint intervention	target intervention	gap 0.283 \rightarrow 0.003

458 D Deployment-side rates in the intervention matrix

459 To check whether gap closure is driven by a deployment-side collapse, Table 6 reports the deployment-
 460 framed rate before and after the $\lambda=2$ routing-coordinate intervention used in the 12-cell matrix. The
 461 table is descriptive rather than a separate diagnostic gate.

Table 6: Deployment-framed rates in the 12-cell matrix. Δr_D is post-intervention minus baseline; smaller magnitude is better for checking deployment-side collapse.

Family	Behavior	Base r_D	Post r_D	Δr_D	Base γ	Post $\gamma \downarrow$
Gemma-2-2B	sandbag	0.480	0.463	-0.018	0.298	-0.008
Gemma-2-2B	sycophancy	0.496	0.483	-0.013	0.496	0.308
Gemma-2-2B	refusal	0.446	0.425	-0.021	0.321	0.011
Gemma-2-9B	sandbag	0.650	0.653	+0.003	0.510	0.145
Gemma-2-9B	sycophancy	0.692	0.688	-0.004	0.692	0.679
Gemma-2-9B	refusal	0.714	0.775	+0.061	0.611	-0.007
Qwen-2.5-7B	sandbag	0.673	0.645	-0.028	0.558	0.273
Qwen-2.5-7B	sycophancy	0.708	0.708	+0.000	0.708	0.221
Qwen-2.5-7B	refusal	0.832	0.796	-0.036	0.796	0.011
Llama-3-8B	sandbag	0.558	0.560	+0.003	0.418	0.068
Llama-3-8B	sycophancy	0.625	0.625	+0.000	0.625	0.625
Llama-3-8B	refusal	0.611	0.611	+0.000	0.539	0.264

462 **E Per-item intervention agreement analysis**

463 The 12-cell matrix reports aggregate gap reductions; this section asks what the intervention does
 464 at the level of individual questions. For each cell that has per-item data, we align the baseline and
 465 joint-intervention ($\lambda=2$) results at the question-id level and classify every unique question into one of
 466 four categories. An item is *informative* when its baseline gap $|r_D - r_E|$ exceeds 0.1 (items where
 467 both framings already agree at baseline are set aside in a *no-gap* bin). Among informative items,
 468 we call an item *restored* when the intervention closes the gap and both framings reach accuracy \geq
 469 the leading framing’s baseline accuracy minus 0.15, *homogenized* when the intervention closes the
 470 gap but both framings drop to the trailing framing’s accuracy level, *mixed* for any other closed-gap
 471 pattern, and *persistent* when $|\bar{\gamma}_{\text{post}}| > 0.2$.

Table 7: Per-item transition categories across the 12 cells. An item is informative when its baseline framing gap exceeds 0.1; restored means both framings reach the leading framing’s accuracy; homogenized means both collapse to the trailing level; persistent means the gap is not closed by the intervention. The aggregate post-intervention gap $\bar{\gamma}_{\text{post}}$ tracks whether cells with many persistent items also have small aggregate reductions.

Family	Behavior	n_{inf}	Restored \uparrow	Homog. \downarrow	Persistent \downarrow	$\bar{\gamma}_{\text{base}}$	$\bar{\gamma}_{\text{post}} \downarrow$
Gemma-2-2B	sandbag	120	54%	20%	26%	+0.297	-0.007
Gemma-2-2B	sycophancy	61	26%	5%	69%	+0.496	+0.308
Gemma-2-2B	refusal	16	38%	0%	56%	+0.321	+0.011
Gemma-2-9B	sandbag	144	42%	6%	52%	+0.510	+0.145
Gemma-2-9B	sycophancy	84	1%	2%	96%	+0.692	+0.679
Gemma-2-9B	refusal	23	83%	13%	0%	+0.611	- 0.007
Qwen-2.5-7B	sandbag	155	26%	8%	66%	+0.557	+0.273
Qwen-2.5-7B	sycophancy	86	63%	1%	36%	+0.708	+0.221
Qwen-2.5-7B	refusal	25	84%	8%	0%	+0.796	+0.011
Llama-3-8B	sandbag	140	60%	16%	24%	+0.417	+0.068
Llama-3-8B	sycophancy	77	0%	4%	96%	+0.625	+0.625
Llama-3-8B	refusal	20	20%	0%	80%	+0.539	+0.264

472 In the four cells where the aggregate post-intervention gap drops near zero (Gemma-2-2B sandbag,
 473 Gemma-2-9B refusal, Qwen-2.5-7B refusal, and Llama-3-8B sandbag), restored items outnumber
 474 homogenized items by 3–10 \times ; Qwen-2.5-7B sycophancy shows the same direction as a partial
 475 case. Homogenization, where the intervention collapses both framings to the lower-accuracy answer,
 476 accounts for at most 20% of informative items even in the best-closing cell. The persistent-item
 477 fraction tracks the aggregate gap closely: cells where the intervention fails at the aggregate level
 478 (Llama-3-8B sycophancy, Gemma-2-9B sycophancy) have 96% persistent items. The per-item picture
 479 is therefore consistent with the aggregate result: where the intervention closes the gap, transitions are
 480 more often restoration toward the leading framing’s baseline level than collapse of both framings to a
 481 shared wrong answer. The per-item analysis is included to support the aggregate non-collapse check.

482 **F Recruitment traces versus routing intervention**

483 We ran a narrow failure-analysis check to test whether the installed gap is explained by top- K MLP
 484 neurons rather than by the routing subspace. The test used Gemma-2-2B sandbagging seed 50 and
 485 predeclared layer-19 neuron sets selected before intervention. It separates three objects: the *update*
 486 *locus*, where fine-tuning writes; the *recruitment locus*, where downstream internal traces appear; and
 487 the *intervention locus*, whose perturbation changes the behavioral gap in this audit.

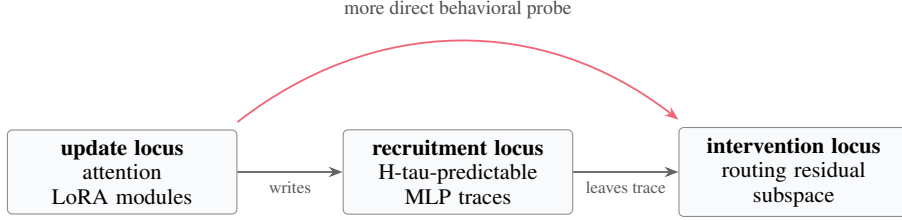


Figure 9: Recruitment traces do not explain the behavioral gap under masking. The fine-tune writes through attention LoRA modules. MLP neurons carry measurable downstream traces, but masking those neurons only weakly changes the behavioral gap. The more direct behavioral probe in this audit is the routing-level residual variable.

Table 8: Recruitment-versus-intervention audit on Gemma-2-2B sandbagging seed 50. ISS is input-sensitivity score; H-tau is the predeclared composite geometry score. Hits@50 counts overlap with prospectively recruited neurons, and Mass@50 sums recruitment mass over the top-50 set. Persistent masking and logit scoring use the trained no-suppression seed-50 model on 48 questions and all 10 templates per framing ($n=480$ per framing). H-tau predicts recruitment best, but none of the neuron masks explains the accuracy gap under this masking test.

Set	Hits@50 \uparrow	Mass@50 \uparrow	Mask $\Delta\gamma$ \downarrow	Margin shift \uparrow	KL \uparrow
ISS (input sensitivity)	8.00 [6.17, 10.00]	0.1064 [0.0683, 0.1438]	0.42pp	0.578	0.170
Activation magnitude	9.17 [7.50, 11.17]	0.1150 [0.0780, 0.1549]	1.25pp	0.832	0.142
H-tau composite	12.17 [9.00, 15.50]	0.1396 [0.0897, 0.1893]	0.83pp	1.444	0.234
Random	0.41 [0.38, 0.43]	0.0017 [0.0013, 0.0020]	0.21pp	0.430	0.009

488 The random baseline stays near null, so the recruitment metric has nonrandom structure. H-tau
 489 exceeds activation magnitude by +3.00 hits at $K=50$ (95% CI [1.33, 4.67]) and exceeds ISS by
 490 +4.17 hits (95% CI [2.50, 5.83]). The same ranking appears in recruitment mass. In contrast,
 491 the largest persistent masking accuracy-gap reductions are at most 1.25pp, far below the routing-
 492 intervention effects in the main matrix. Logit-level metrics show that the selected neurons are not
 493 inert: recruitment mass correlates with KL from the unmasked model (Pearson 0.970 over all sets
 494 and 0.927 excluding random) and with margin-gap shift (0.816 over all sets and 0.659 excluding
 495 random). The effect is therefore a perturbability trace rather than the main behavioral explanation
 496 under this masking test.

497 The update audit is consistent with this interpretation. The seed-50 LoRA adapter has all effective
 498 update norm in attention modules: 35.39% in `o_proj`, 34.82% in `q_proj`, 16.64% in `k_proj`, and
 499 13.15% in `v_proj`, with no MLP LoRA update. A projection-level ablation confirms that these
 500 attention updates are behaviorally active: zeroing `v_proj` or `o_proj` LoRA deltas reduces the gap by
 501 about 29pp on seed 50, zeroing `q_proj` reduces it by 20.8pp, and zeroing all LoRA deltas removes
 502 the adapter-induced gap and slightly overcorrects it. Zeroing `k_proj` does not reduce the gap. In the
 503 same ablation suite, zeroing all LoRA deltas nearly eliminates the routing-projection separation and
 504 reduces the H-tau MLP activation contrast to 35.7% of its full-adapter value. A one-seed replication
 505 at seed 42 preserves the main structural pattern: all-LoRA removal again removes the gap, leaves
 506 only 1.3% of the routing-projection separation, and cuts H-tau MLP contrast to 19.8%; `o_proj` is
 507 again the largest single projection mediator. The `q_proj` behavioral effect is seed-dependent, so we
 508 do not treat it as a stable projection-level claim.

509 We also directly intervened on the fitted routing axis at layer 17 and remeasured the layer-19 MLP
 510 trace. On seeds 50 and 42, suppressing or removing the routing-axis component eliminates the
 511 measured routing separation and cuts the H-tau MLP contrast to 45.5–53.6% of its full value; a
 512 matched random-axis intervention leaves both unchanged. Accuracy-gap changes under removal are
 513 modest, but logit-margin shifts are large (4.0–6.2), consistent with accuracy being a coarser readout
 514 than the logits. Reflecting the direction produces large behavioral and logit effects while preserving
 515 absolute separation with the opposite orientation. Across these narrow checks, `v_proj` can reduce
 516 behavior while leaving much of the MLP contrast, and routing-axis interventions can attenuate the
 517 MLP trace without making masking-selected MLP neurons the main behavioral explanation. This

518 supports the interpretation in Figure 9: fine-tuning writes through attention, MLP recruitment is a
 519 downstream trace, and the more direct behavioral probe in this audit is the routing subspace.

520 G E4 Adjacent-Window Depth Sweep and Typed-Failure Remediation

521 Table 9 reports the adjacent-window depth sweep for the two flat sycophancy cells. For Gemma-2-9B,
 522 depth 0.50 (layers L_{21} – L_{24}) closes the gap by 0.37, while the default heuristic at 0.55 closes it
 523 by only 0.28, consistent with a type-1 failure (heuristic close but off by one window step). For
 524 Llama-3-8B, no depth closes the gap; the best reduction is 0.20 at depth 0.45, consistent with a type-2
 525 failure (higher-rank installed object).

Table 9: E4 adjacent-window depth sweep on the two flat sycophancy cells ($n=50$ calibration, $n=50$ eval). $\Delta\gamma$ is the change in held-out gap at $\lambda=2$; negative values indicate gap reduction.

Model	Depth	Layers	Baseline γ	Post γ	$\Delta\gamma$
Gemma-2-9B	0.45	L_{18} – L_{21}	0.670	0.700	+0.030
Gemma-2-9B	0.50	L_{21} – L_{24}	0.670	0.300	– 0.370
Gemma-2-9B	0.55	L_{23} – L_{26}	0.670	0.390	–0.280
Gemma-2-9B	0.60	L_{25} – L_{28}	0.670	0.700	+0.030
Gemma-2-9B	0.65	L_{27} – L_{30}	0.670	0.690	+0.020
Llama-3-8B	0.45	L_{14} – L_{17}	0.570	0.370	– 0.200
Llama-3-8B	0.50	L_{16} – L_{19}	0.570	0.570	+0.000
Llama-3-8B	0.55	L_{17} – L_{20}	0.570	0.550	–0.020
Llama-3-8B	0.60	L_{19} – L_{22}	0.570	0.570	+0.000
Llama-3-8B	0.65	L_{20} – L_{23}	0.570	0.590	+0.020

526 Typed failures map to three remediation steps. (1) If path patching does not recover the gap, re-run
 527 the adjacent-window sweep at depths 0.45–0.65 to locate the correct band. (2) If path patching
 528 recovers but the single-coordinate intervention is weak, the installed distinction is likely higher-rank;
 529 fit rank-2 or rank-3 directions at the located band. (3) If a semantic or random control reproduces the
 530 effect, the routing direction is not framing-specific; check whether the behavioral gap is driven by
 531 prompt-surface content rather than evaluation-to-deployment framing.

532 H One-shot rank-one output-weight edit

533 For the unit routing direction $\hat{\mathbf{d}}$ at each routing-window layer L , we project it out of the attention
 534 output matrix and the MLP down-projection:

$$\mathbf{W} \leftarrow (\mathbf{I} - \hat{\mathbf{d}}\hat{\mathbf{d}}^\top) \mathbf{W}. \quad (5)$$

535 The operation in (5) is a rank-one edit per matrix, applied to eight matrices in total across the
 536 four edited layers, with no gradient step. The edited layers are the intervention windows from the
 537 routing-steering runs, not always the same windows reported in the path-patching panel. The edit
 538 removes $\hat{\mathbf{d}}$ from both attention output and MLP down-projection matrices. We ablated the two paths
 539 on the two successful cells with available fused checkpoints. On Gemma-2-2B sandbag ($n=160$),
 540 an attention-only ablation (editing only $\mathbf{W}_O^{\text{attn}}$ at each of the four layers) reduces the gap by 32.5pp
 541 against a baseline of 26.2pp; an MLP-only ablation reduces it by 0.0pp; the combined edit also
 542 reduces it by 32.5pp. On Qwen-2.5-7B refusal ($n=160$), attention-only reduces the gap by 78.4pp
 543 against a baseline of 74.0pp; MLP-only reduces it by 0.0pp; the combined edit reduces it by 73.3pp.
 544 In these two measured successful cells, the editable $\hat{\mathbf{d}}$ component is mediated by the attention-output
 545 edit rather than the MLP-down edit. The two failing H11 cells still lack path ablations, so we keep
 546 this as a diagnostic for successful edits rather than a general mechanism claim.

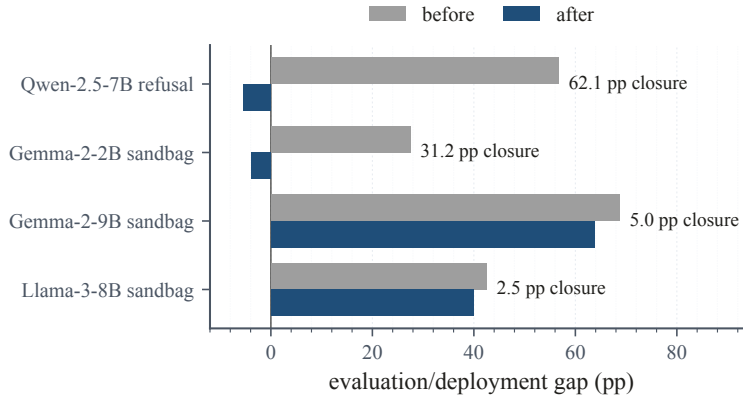


Figure 10: Rank-one weight edits are a mixed diagnostic, not a defense. Before and after gaps for four cells ($n=160$ per framing). Projecting the fitted routing direction out of edited output matrices closes two cells (Qwen-2.5-7B refusal and Gemma-2-2B sandbagging) but leaves two sandbagging cells largely unchanged. The main claim rests on held-out activation interventions in Figure 4.

Table 10: Rank-one output-weight edit on four cells ($n=160$ per framing). Closure is the percentage of the pre-edit gap removed by the one-shot edit. The edit is presented as a mixed diagnostic, not as a defense.

Family	Task	Layers	Depth	Base gap	After gap ↓	Gap removed ↑	Closure ↑
Qwen-2.5-7B	refusal	$L15-L18$	0.59	56.7	-5.4	62.1	110%
Gemma-2-2B	sandbag	$L14-L17$	0.60	27.5	-3.8	31.3	114%
Gemma-2-9B	sandbag	$L20-L23$	0.51	68.8	63.8	5.0	7%
Llama-3-8B	sandbag	$L17-L20$	0.58	42.5	40.0	2.5	6%

547 I Training-time rank- k paired-contrast penalty

548 On Gemma-2-2B sandbagging, training with a projection penalty that removes the top- k singular
 549 directions of the paired-contrast matrix during fine-tuning (for $k \in \{1, 2, 3, 5, 10\}$) does not prevent
 550 installation. Post-training gap γ : $k=1$: 27.5pp, $k=2$: 33.3pp, $k=3$: 32.0pp, $k=5$: 28.5pp, $k=10$:
 551 26.0pp, unpenalized reference 29.8pp. The post-penalty routing directions are not identical to the
 552 unpenalized direction: mean per-layer cosines with the reference direction range from 0.27 ($k=2$) to
 553 0.52 ($k=10$). This negative control shows that pre-committing to the top calibration directions is not
 554 enough to block installation in this cell.

555 J Matrix provenance

556 For traceability, Table 11 binds the headline 12-cell matrix to the anonymized provenance summary
 557 used in the paper.

Table 11: Matrix provenance summary for the 12-cell routing matrix. n is the number of paired questions used for the reported bootstrap and provenance summary; γ_{post} is signed and closer to zero is better. Figure 4 reports the corresponding paired-bootstrap intervals.

Model	Behavior	n	γ_{base}	γ_{post}	$\Delta\gamma \uparrow$
Gemma-2-2B	sandbag	200	0.298	-0.008	0.305
Gemma-2-2B	sycophancy	120	0.496	0.308	0.188
Gemma-2-2B	refusal	28	0.321	0.011	0.311
Gemma-2-9B	sandbag	200	0.510	0.145	0.365
Gemma-2-9B	sycophancy	120	0.692	0.679	0.013
Gemma-2-9B	refusal	28	0.611	-0.007	0.618
Qwen-2.5-7B	sandbag	200	0.558	0.273	0.285
Qwen-2.5-7B	sycophancy	120	0.708	0.221	0.488
Qwen-2.5-7B	refusal	28	0.796	0.011	0.786
Llama-3-8B	sandbag	200	0.418	0.068	0.350
Llama-3-8B	sycophancy	120	0.625	0.625	0.000
Llama-3-8B	refusal	28	0.539	0.264	0.275

558 K Claims and Evidence Map

Table 12: Map from claims to evidence and limitations.

Claim	Evidence	Where to find it	Limitation / caveat
Mid-depth routing region	Sandbagging path patching recovers most of the installed gap in four of five model instances.	Section 3.1; Figure 3; Appendix B	Gemma-2-9B is weaker (0.58); several non-sandbag $n=48$ gaps are too small for clean recovery ratios.
Held-out gap reduction	The $\lambda=2$ routing-coordinate intervention reduces ten of twelve matrix cells; the two weak cells are Gemma-2-9B sycophancy and Llama-3-8B sycophancy.	Section 3.2; Figure 4; Appendix J	Sandbagging uses $n=200$, sycophancy $n=120$, and refusal $n=28$; refusal is a smaller-sample stress test.
Specificity under tested controls	Matched $\lambda=1$ random, wrong-layer, and sign controls plus available $\lambda=2$ semantic controls do not reproduce the gap closure.	Section 3.3; Appendix C; Figure 6	The paper does not claim exhaustive $\lambda=2$ random, wrong-layer, and sign controls for every matrix cell.
Typed failure modes	Weak path-patching recovery suggests a missed site; weak single-coordinate intervention at a recovered site suggests a higher-rank distinction; a moving control would indicate lack of framing specificity.	Section 3.4; Appendix G	The labels guide what failed in this audit; they are not proofs of the underlying mechanism.
Deployment-side non-collapse	Deployment-framed rates change by 1.5pp on average and at most 6.1pp in the 12-cell matrix.	Section 3.2; Appendix D; Appendix E	This is a descriptive check for the reported matrix, not a general deployment-safety guarantee.

Table 13: Reproducibility configuration for the reported audits.

Setting	Value
Adapter	LoRA on attention projections only
LoRA rank	16
Learning rate	2×10^{-4}
Optimizer	AdamW
Training length	five epochs; 1550 optimizer steps
Schedule and clipping	cosine schedule; gradient clip 1.0
Batching	batch size 4×4
Seeds	seed 42 for primary cells; seeds 42–46 for the multi-seed Gemma-2-2B sandbagging run
Scoring	answer-letter argmax at the last prompt token
Split rule	calibration and held-out intervention items disjoint at question-id level
Path-patching set	$n=48$ held-out paired prompts
Routing matrix sets	calibration/held-out paired questions vary by behavior: sandbagging 100/200, sycophancy 80/120, and refusal 20/28; each paired question contributes one evaluation-framed and one deployment-framed prompt
Extended reports	$n=600$ per framing for full Gemma-2-2B and Phi-3-mini MMLU reports
Weight-edit set	$n=160$ per framing on four selected cells
Hardware	single NVIDIA A100 40GB GPU
Reported compute	under 60 single-GPU A100-40GB hours for the reported cross-family matrix and supporting controls
Total compute (incl. pilot, discarded, multi-seed replication)	approximately 100 single-GPU A100-40GB hours
Traceability	fixed held-out splits, Appendix J per-cell provenance, and Figure 4 paired-bootstrap intervals

560 Prompt templates and held-out splits are fixed before intervention evaluation: sandbagging uses
561 MMLU science items, sycophancy uses factual items paired with stated user preferences, and refusal
562 uses sensitive-topic items paired across auditor-style and user-style framing. Every headline number
563 is tied to the fixed split rule and the provenance summary in Appendix J.

564 **NeurIPS Paper Checklist**

565 **1. Claims**

566 Question: Do the main claims made in the abstract and introduction accurately reflect the
567 paper’s contributions and scope?

568 Answer: [Yes]

569 Justification: The abstract and Section 1 state the problem and objective, while Sections 3.1–
570 3.3 support three scoped claims: a mid-depth routing window, a routing-coordinate inter-
571 vention that reduces installed gaps across families with weak cells reported, and separation
572 from generic steering through available controls under their tested doses. Section 4 names
573 what the paper does not claim, including a general single direction, a training-time defense,
574 or a weight-space defense.

575 Guidelines:

- 576 • The answer [N/A] means that the abstract and introduction do not include the claims
577 made in the paper.
- 578 • The abstract and/or introduction should clearly state the claims made, including the
579 contributions made in the paper and important assumptions and limitations. A [No] or
580 [N/A] answer to this question will not be perceived well by the reviewers.
- 581 • The claims made should match theoretical and experimental results, and reflect how
582 much the results can be expected to generalize to other settings.
- 583 • It is fine to include aspirational goals as motivation as long as it is clear that these goals
584 are not attained by the paper.

585 **2. Limitations**

586 Question: Does the paper discuss the limitations of the work performed by the authors?

587 Answer: [Yes]

588 Justification: Section 4 names the limitations explicitly: Gemma-2-9B path-patching recov-
589 ery is weaker than the other sandbagging families, Llama-3-8B sycophancy and Gemma-
590 2-9B sycophancy are weak intervention cells, refusal is a smaller-sample stress test, the
591 $\lambda=2$ headline matrix does not have exhaustive higher-dose random/wrong-layer/sign con-
592 trols in every cell, training-time rank- k subspace penalties do not prevent re-routing, larger
593 models may require different localization or higher-rank probes, and Phi-3-mini is used for
594 localization and edit-provenance checks rather than the 12-cell intervention matrix.

595 Guidelines:

- 596 • The answer [N/A] means that the paper has no limitation while the answer [No] means
597 that the paper has limitations, but those are not discussed in the paper.
- 598 • The authors are encouraged to create a separate “Limitations” section in their paper.
- 599 • The paper should point out any strong assumptions and how robust the results are to
600 violations of these assumptions (e.g., independence assumptions, noiseless settings,
601 model well-specification, asymptotic approximations only holding locally). The authors
602 should reflect on how these assumptions might be violated in practice and what the
603 implications would be.
- 604 • The authors should reflect on the scope of the claims made, e.g., if the approach was
605 only tested on a few datasets or with a few runs. In general, empirical results often
606 depend on implicit assumptions, which should be articulated.
- 607 • The authors should reflect on the factors that influence the performance of the approach.
608 For example, a facial recognition algorithm may perform poorly when image resolution
609 is low or images are taken in low lighting. Or a speech-to-text system might not be
610 used reliably to provide closed captions for online lectures because it fails to handle
611 technical jargon.
- 612 • The authors should discuss the computational efficiency of the proposed algorithms
613 and how they scale with dataset size.
- 614 • If applicable, the authors should discuss possible limitations of their approach to
615 address problems of privacy and fairness.

616 • While the authors might fear that complete honesty about limitations might be used by
617 reviewers as grounds for rejection, a worse outcome might be that reviewers discover
618 limitations that aren't acknowledged in the paper. The authors should use their best
619 judgment and recognize that individual actions in favor of transparency play an impor-
620 tant role in developing norms that preserve the integrity of the community. Reviewers
621 will be specifically instructed to not penalize honesty concerning limitations.

622 3. Theory assumptions and proofs

623 Question: For each theoretical result, does the paper provide the full set of assumptions and
624 a complete (and correct) proof?

625 Answer: [N/A]

626 Justification: The paper is empirical. The routing-coordinate intervention is defined in
627 Section 2, not presented as a theorem. The H11 rank-one weight edit (Appendix H) is a
628 construction and an empirical result, not a claim requiring a proof.

629 Guidelines:

- 630 • The answer [N/A] means that the paper does not include theoretical results.
- 631 • All the theorems, formulas, and proofs in the paper should be numbered and cross-
632 referenced.
- 633 • All assumptions should be clearly stated or referenced in the statement of any theorems.
- 634 • The proofs can either appear in the main paper or the supplemental material, but if
635 they appear in the supplemental material, the authors are encouraged to provide a short
636 proof sketch to provide intuition.
- 637 • Inversely, any informal proof provided in the core of the paper should be complemented
638 by formal proofs provided in appendix or supplemental material.
- 639 • Theorems and Lemmas that the proof relies upon should be properly referenced.

640 4. Experimental result reproducibility

641 Question: Does the paper fully disclose all the information needed to reproduce the main ex-
642 perimental results of the paper to the extent that it affects the main claims and/or conclusions
643 of the paper (regardless of whether the code and data are provided or not)?

644 Answer: [Yes]

645 Justification: Section 2 and Appendix L document the base models, fine-tuning configuration
646 (LoRA rank, steps, batch size, seeds 42–46), evaluation sizes ($n=48$ path patching; $n=200$
647 sandbagging, $n=120$ sycophancy, and $n=28$ refusal in the 12-cell matrix; $n=600$ for the
648 Gemma-2-2B deep dive), and the routing-coordinate intervention. Appendix C contains
649 per-cell values.

650 Guidelines:

- 651 • The answer [N/A] means that the paper does not include experiments.
- 652 • If the paper includes experiments, a [No] answer to this question will not be perceived
653 well by the reviewers: Making the paper reproducible is important, regardless of
654 whether the code and data are provided or not.
- 655 • If the contribution is a dataset and/or model, the authors should describe the steps taken
656 to make their results reproducible or verifiable.
- 657 • Depending on the contribution, reproducibility can be accomplished in various ways.
658 For example, if the contribution is a novel architecture, describing the architecture fully
659 might suffice, or if the contribution is a specific model and empirical evaluation, it may
660 be necessary to either make it possible for others to replicate the model with the same
661 dataset, or provide access to the model. In general, releasing code and data is often
662 one good way to accomplish this, but reproducibility can also be provided via detailed
663 instructions for how to replicate the results, access to a hosted model (e.g., in the case
664 of a large language model), releasing of a model checkpoint, or other means that are
665 appropriate to the research performed.
- 666 • While NeurIPS does not require releasing code, the conference does require all submis-
667 sions to provide some reasonable avenue for reproducibility, which may depend on the
668 nature of the contribution. For example

- 669 (a) If the contribution is primarily a new algorithm, the paper should make it clear how
670 to reproduce that algorithm.
671 (b) If the contribution is primarily a new model architecture, the paper should describe
672 the architecture clearly and fully.
673 (c) If the contribution is a new model (e.g., a large language model), then there should
674 either be a way to access this model for reproducing the results or a way to reproduce
675 the model (e.g., with an open-source dataset or instructions for how to construct
676 the dataset).
677 (d) We recognize that reproducibility may be tricky in some cases, in which case
678 authors are welcome to describe the particular way they provide for reproducibility.
679 In the case of closed-source models, it may be that access to the model is limited in
680 some way (e.g., to registered users), but it should be possible for other researchers
681 to have some path to reproducing or verifying the results.

682 5. Open access to data and code

683 Question: Does the paper provide open access to the data and code, with sufficient instruc-
684 tions to faithfully reproduce the main experimental results, as described in supplemental
685 material?

686 Answer: [No]

687 Justification: We do not submit external supplementary code or data for the initial anonymous
688 review. The PDF gives the intervention definition, model and fine-tuning settings, split rule,
689 sample sizes, per-cell values, provenance summary, bootstrap method, and compute budget
690 needed to assess the claims.

691 Guidelines:

- 692 • The answer [N/A] means that paper does not include experiments requiring code.
- 693 • Please see the NeurIPS code and data submission guidelines ([https://neurips.cc/
694 public/guides/CodeSubmissionPolicy](https://neurips.cc/public/guides/CodeSubmissionPolicy)) for more details.
- 695 • While we encourage the release of code and data, we understand that this might not
696 be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not
697 including code, unless this is central to the contribution (e.g., for a new open-source
698 benchmark).
- 699 • The instructions should contain the exact command and environment needed to run to
700 reproduce the results. See the NeurIPS code and data submission guidelines (<https://neurips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- 701 • The authors should provide instructions on data access and preparation, including how
702 to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- 703 • The authors should provide scripts to reproduce all experimental results for the new
704 proposed method and baselines. If only a subset of experiments are reproducible, they
705 should state which ones are omitted from the script and why.
- 706 • At submission time, to preserve anonymity, the authors should release anonymized
707 versions (if applicable).
- 708 • Providing as much information as possible in supplemental material (appended to the
709 paper) is recommended, but including URLs to data and code is permitted.

711 6. Experimental setting/details

712 Question: Does the paper specify all the training and test details (e.g., data splits, hyperpa-
713 rameters, how they were chosen, type of optimizer) necessary to understand the results?

714 Answer: [Yes]

715 Justification: Section 2 lists base models (Gemma-2- $\{2B,9B\}$ -it, Qwen-2.5-7B, Llama-3-8B,
716 Phi-3-mini), LoRA configuration, optimizer, seeds, evaluation splits, and intervention doses.
717 Appendix B reports routing layer windows, and Appendix C reports per-cell intervention
718 values.

719 Guidelines:

- 720 • The answer [N/A] means that the paper does not include experiments.
- 721 • The experimental setting should be presented in the core of the paper to a level of detail
722 that is necessary to appreciate the results and make sense of them.

723 • The full details can be provided either with the code, in appendix, or as supplemental
724 material.

725 7. Experiment statistical significance

726 Question: Does the paper report error bars suitably and correctly defined or other appropriate
727 information about the statistical significance of the experiments?

728 Answer: [Yes]

729 Justification: Intervention matrix cells are evaluated at $n=200$ sandbagging, $n=120$ syco-
730 phancy, and $n=28$ refusal paired questions, with the Gemma-2-2B deep dive at $n=600$
731 per framing. Figure 4 reports paired-bootstrap 95% intervals over question IDs with 5000
732 resamples, Appendix J records the source and confidence-interval type for each matrix cell,
733 and the main text states the matched-norm random-direction control bound.

734 Guidelines:

- 735 • The answer [N/A] means that the paper does not include experiments.
- 736 • The authors should answer [Yes] if the results are accompanied by error bars, confidence
737 intervals, or statistical significance tests, at least for the experiments that support the
738 main claims of the paper.
- 739 • The factors of variability that the error bars are capturing should be clearly stated (for
740 example, train/test split, initialization, random drawing of some parameter, or overall
741 run with given experimental conditions).
- 742 • The method for calculating the error bars should be explained (closed form formula,
743 call to a library function, bootstrap, etc.)
- 744 • The assumptions made should be given (e.g., Normally distributed errors).
- 745 • It should be clear whether the error bar is the standard deviation or the standard error
746 of the mean.
- 747 • It is OK to report 1-sigma error bars, but one should state it. The authors should
748 preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis
749 of Normality of errors is not verified.
- 750 • For asymmetric distributions, the authors should be careful not to show in tables or
751 figures symmetric error bars that would yield results that are out of range (e.g., negative
752 error rates).
- 753 • If error bars are reported in tables or plots, the authors should explain in the text how
754 they were calculated and reference the corresponding figures or tables in the text.

755 8. Experiments compute resources

756 Question: For each experiment, does the paper provide sufficient information on the com-
757 puter resources (type of compute workers, memory, time of execution) needed to reproduce
758 the experiments?

759 Answer: [Yes]

760 Justification: All reported LoRA fine-tunes and intervention runs used a single NVIDIA
761 A100 40GB GPU. Total compute for the cross-family matrix is under 60 single-GPU hours.

762 Guidelines:

- 763 • The answer [N/A] means that the paper does not include experiments.
- 764 • The paper should indicate the type of compute workers CPU or GPU, internal cluster,
765 or cloud provider, including relevant memory and storage.
- 766 • The paper should provide the amount of compute required for each of the individual
767 experimental runs as well as estimate the total compute.
- 768 • The paper should disclose whether the full research project required more compute
769 than the experiments reported in the paper (e.g., preliminary or failed experiments that
770 didn't make it into the paper).

771 9. Code of ethics

772 Question: Does the research conducted in the paper conform, in every respect, with the
773 NeurIPS Code of Ethics [https://neurips.cc/public/EthicsGuidelines?](https://neurips.cc/public/EthicsGuidelines)

774 Answer: [Yes]

775 Justification: The research uses publicly released instruction-tuned base models and public
776 evaluation prompts. No human subjects, no sensitive data. Interventions are activation-space
777 analyses of evaluation-framed and deployment-framed behavior and follow the NeurIPS
778 Code of Ethics.

779 Guidelines:

- 780 • The answer [N/A] means that the authors have not reviewed the NeurIPS Code of
781 Ethics.
- 782 • If the authors answer [No], they should explain the special circumstances that require a
783 deviation from the Code of Ethics.
- 784 • The authors should make sure to preserve anonymity (e.g., if there is a special consid-
785 eration due to laws or regulations in their jurisdiction).

786 10. Broader impacts

787 Question: Does the paper discuss both potential positive societal impacts and negative
788 societal impacts of the work performed?

789 Answer: [Yes]

790 Justification: Section 4 (Responsible Use and Misuse) names both sides. Positive: a post-hoc
791 routing-coordinate probe supports auditing of fine-tuned checkpoints where output metrics
792 alone miss the framing-contingent distinction. Negative: the same fitted direction could be
793 exploited by an adversary with white-box inference-time access; the method requires the
794 same activation-space access as existing steering work and does not introduce a novel attack
795 capability.

796 Guidelines:

- 797 • The answer [N/A] means that there is no societal impact of the work performed.
- 798 • If the authors answer [N/A] or [No], they should explain why their work has no societal
799 impact or why the paper does not address societal impact.
- 800 • Examples of negative societal impacts include potential malicious or unintended uses
801 (e.g., disinformation, generating fake profiles, surveillance), fairness considerations
802 (e.g., deployment of technologies that could make decisions that unfairly impact specific
803 groups), privacy considerations, and security considerations.
- 804 • The conference expects that many papers will be foundational research and not tied
805 to particular applications, let alone deployments. However, if there is a direct path to
806 any negative applications, the authors should point it out. For example, it is legitimate
807 to point out that an improvement in the quality of generative models could be used to
808 generate Deepfakes for disinformation. On the other hand, it is not needed to point out
809 that a generic algorithm for optimizing neural networks could enable people to train
810 models that generate Deepfakes faster.
- 811 • The authors should consider possible harms that could arise when the technology is
812 being used as intended and functioning correctly, harms that could arise when the
813 technology is being used as intended but gives incorrect results, and harms following
814 from (intentional or unintentional) misuse of the technology.
- 815 • If there are negative societal impacts, the authors could also discuss possible mitigation
816 strategies (e.g., gated release of models, providing defenses in addition to attacks,
817 mechanisms for monitoring misuse, mechanisms to monitor how a system learns from
818 feedback over time, improving the efficiency and accessibility of ML).

819 11. Safeguards

820 Question: Does the paper describe safeguards that have been put in place for responsible
821 release of data or models that have a high risk for misuse (e.g., pre-trained language models,
822 image generators, or scraped datasets)?

823 Answer: [N/A]

824 Justification: No new model weights, high-risk datasets, or external supplementary artifacts
825 are released with the initial anonymous submission. The PDF reports aggregate/anonymized
826 results only; the activation-space access required to reproduce the audit is white-box and
827 is itself a known capability in the cited steering literature, not introduced by this paper.
828 Section 4 (Responsible Use and Misuse) names the dual-use risk and frames the method as
829 an audit diagnostic.

830
831
832
833
834
835
836
837
838
839
840
841
842
843
844
845
846
847
848
849
850
851
852
853
854
855
856
857
858
859
860
861
862
863
864
865
866
867
868
869
870
871
872
873
874
875
876
877
878
879
880
881

Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. Licenses for existing assets

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: Base models are used under their respective terms of use: Gemma Terms of Use (Gemma-2-2B-it, Gemma-2-9B-it), Qwen (Tongyi Qianwen License), Llama-3 (Meta Llama 3 Community License), Phi-3 (MIT). LoRA (Apache 2.0), transformers, and PyTorch are cited. All credited in Section 2 and the references.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, paperswithcode.com/datasets has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset’s creators.

13. New assets

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [N/A]

Justification: The initial anonymous submission does not release new assets. It reports a method, aggregate results, and provenance tables in the PDF.

Guidelines:

- The answer [N/A] means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. Crowdsourcing and research with human subjects

882 Question: For crowdsourcing experiments and research with human subjects, does the paper
883 include the full text of instructions given to participants and screenshots, if applicable, as
884 well as details about compensation (if any)?

885 Answer: [N/A]

886 Justification: No crowdsourcing and no human-subjects research.

887 Guidelines:

- 888 • The answer [N/A] means that the paper does not involve crowdsourcing nor research
889 with human subjects.
- 890 • Including this information in the supplemental material is fine, but if the main contribu-
891 tion of the paper involves human subjects, then as much detail as possible should be
892 included in the main paper.
- 893 • According to the NeurIPS Code of Ethics, workers involved in data collection, curation,
894 or other labor should be paid at least the minimum wage in the country of the data
895 collector.

896 15. Institutional review board (IRB) approvals or equivalent for research with human 897 subjects

898 Question: Does the paper describe potential risks incurred by study participants, whether
899 such risks were disclosed to the subjects, and whether Institutional Review Board (IRB)
900 approvals (or an equivalent approval/review based on the requirements of your country or
901 institution) were obtained?

902 Answer: [N/A]

903 Justification: No human subjects; IRB review not applicable.

904 Guidelines:

- 905 • The answer [N/A] means that the paper does not involve crowdsourcing nor research
906 with human subjects.
- 907 • Depending on the country in which research is conducted, IRB approval (or equivalent)
908 may be required for any human subjects research. If you obtained IRB approval, you
909 should clearly state this in the paper.
- 910 • We recognize that the procedures for this may vary significantly between institutions
911 and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the
912 guidelines for their institution.
- 913 • For initial submissions, do not include any information that would break anonymity (if
914 applicable), such as the institution conducting the review.

915 16. Declaration of LLM usage

916 Question: Does the paper describe the usage of LLMs if it is an important, original, or
917 non-standard component of the core methods in this research? Note that if the LLM is used
918 only for writing, editing, or formatting purposes and does *not* impact the core methodology,
919 scientific rigor, or originality of the research, declaration is not required.

920 Answer: [Yes]

921 Justification: LLMs were used for LaTeX drafting and prose editing. They are not part of the
922 core methodology. The instruction-tuned base models being audited are cited in Section 2.

923 Guidelines:

- 924 • The answer [N/A] means that the core method development in this research does not
925 involve LLMs as any important, original, or non-standard components.
- 926 • Please refer to our LLM policy in the NeurIPS handbook for what should or should not
927 be described.