Broken Symmetry in LLM Refusal: Answer Release Is More Local Than Refusal Restoration

Anonymous ACL submission

Abstract

When a language model refuses to answer a prompt, it is unclear whether the correct answer is erased from its internal representations, or merely suppressed at the output layer. We investigate this mechanism using a controlled withhold setting, which yields perfectly matched answering and refusal trajectories for bidirectional activation patching. We uncover a causal asymmetry, namely a **broken symmetry**, in how refusal operates. Even when a model generates a clean refusal, the correct answer remains linearly recoverable from its hidden states. Furthermore, releasing this withheld answer is a highly local operation, requiring only a single-position patch. Conversely, the reverse operation is not equally local: reimposing suppression requires broader interventions across multiple positions, and assembling a coherent refusal sequence is more difficult still. We further demonstrate that while an average answer-to-refusal displacement vector marks the geometric difference between these states, it fails to act as a reliable, reversible linear control toggle between behaviours. Taken together, our findings show that refusal does not function as a simple symmetric switch. For safety and auditing, this implies that probe recoverability can overestimate true behavioural control, and locating refusal-relevant directions does not reliably grant the ability to steer a model from answering to coherent refusal.

1 Introduction

Suppose a model has enough information to identify a correct answer, but has been instructed not to reveal it. If refusal operates like a simple, local switch within the model's internal states, then turning refusal off should expose the hidden answer, and turning it back on should require a comparable, equally local edit. Recent work makes this switch-like picture compelling: compact activation directions can steer refusal behaviour (Arditi et al., 2024), and contrastive activation-addition provides a standard recipe for intervening on these directions (Zou et al., 2023a; Rimsky et al., 2024).

However, directional steerability does not establish causal symmetry. A geometric direction may mark where answering and refusing states differ, but it does not prove that answer release and full refusal assembly are controlled by matched, equally local inverse edits. This distinction is critical because refusal is not a monolithic behaviour. Suppressing the local decision space (the answer channel) is mechanistically distinct from generating an explicit, stable refusal string (full refusal assembly).

Motivated by this gap, we test the hypothesis that refusal interventions are symmetric: if a highly local edit releases a withheld answer, the matched reverse intervention should locally reimpose suppression. We investigate this across state-of-the-art open-weights models (including the Qwen, Llama, Mistral, and Gemma families). By using a controlled withheld setting, where instruction-forced prompts demand the model hide the correct option to an answerable A/B question, we isolate perfectly matched answering and refusing trajectories with a shared answer-channel readout. we establish our high-resolution mechanistic results mainly on Qwen-family checkpoints, and then provide cross-family support for the same qualitative pattern.

Our experiments reveal that refusal is fundamentally asymmetric. At a high level, the picture is clear: ***the withheld answer remains linearly recoverable, releasing it is a highly local operation, but restoring refusal is not***. Bidirectional patching demonstrates that a single-position edit is sufficient to restore the hidden answer on the answer-channel readout. In contrast, reimposing suppression requires broader patches across multiple positions, and assembling full refusal behaviour is harder still. We call this pattern **broken symmetry**. Furthermore, while we can isolate an average answer-to-

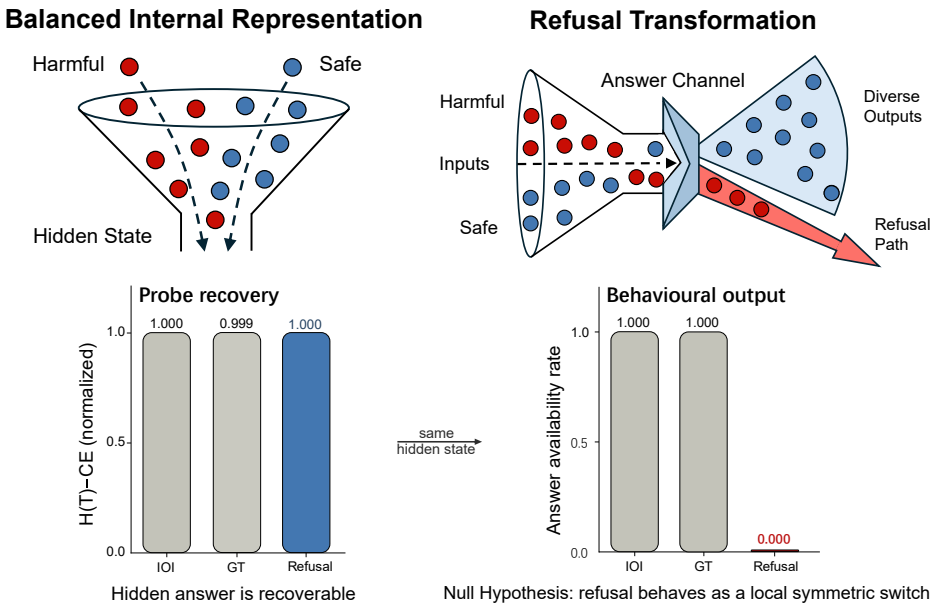

Figure 1: Conceptual picture of hidden-answer suppression and the symmetry hypothesis. A model may internally encode the correct answer while producing a refusal output. If refusal behaves like a local switch, then enabling and disabling it should require comparable local interventions.

refusal displacement vector, adding or removing this shared direction does not act as a reliable, reversible toggle between answering and coherent refusal.

With this distinction established, our paper makes three primary claims:

- **Refusal interventions are not equally local.** Releasing a withheld answer requires only a local, single-position edit, but reimposing answer-channel suppression requires broader, distributed support.
- **Suppression and assembly dissociate.** Suppressing the correct answer channel is causally distinct from assembling a coherent refusal string under intervention.
- **Hidden recoverability overestimates behavioural control.** Because the correct answer remains recoverable even during a clean refusal, probe-only safety audits overestimate true behavioural control.

Across extensive validations over model sizes (from 7B to 32B) and model families (Qwen, Llama, Mistral, and Gemma), the qualitative hidden-state asymmetry holds, while generation-level control and shared-direction interventions show checkpoint-dependent strength. This pattern is consistent with a dissociation between concentrated refusal geometry and reliable behavioural control. Taken together, these results suggest that locating refusal-relevant geometry does not inherently grant causal control over refusal behaviour in LLMs.

2 Related Work

Causal patching and decoder-restricted recovery. Activation patching descends from causal mediation and tracing (Vig et al., 2020; Meng et al., 2022), with recent guidance on metric choice and saturation (Zhang and Nanda, 2024; Heimersheim and Nanda, 2024). On the decoder side, V-information and MDL-style probing treat the decoder family as part of the measured quantity (Xu et al., 2020; Hewitt et al., 2021; Pimentel et al., 2020; Voita and Titov, 2020), motivating our use of progressively richer decoder families—linear probes, MLP probes, affine lenses, and tuned lenses (Belrose et al., 2023; Pal et al., 2023)—to separate recoverability from behavioural control.

Hidden knowledge and refusal geometry. Models can internally encode information that their outputs suppress (Burns et al., 2023; Azaria and Mitchell, 2023; Gekhman et al., 2025; Fedzechkina et al., 2026). In the refusal setting, Arditi et al. (2024) show that a shared refusal direction causally changes behaviour, and Wollschläger et al. (2025) extend this to multi-dimensional concept cones. These works characterise refusal-relevant geometry

but do not test whether answer release and refusal restoration are mirror-local inverse operations.

Activation steering and safety mechanisms. Contrastive activation addition (Zou et al., 2023a; Rimsky et al., 2024) and related work localise safety-relevant features via sparse autoencoders (Yeo et al., 2025), over-refusal mitigation (Dabas et al., 2025), and weight pruning (Orgad et al., 2026). These methods demonstrate directional steerability but do not ask whether a refusal-relevant direction is sufficient for full behaviour assembly. We address this with matched bidirectional patching.

3 Experimental Setup

Every experiment in our pipeline asks two linked questions on the same hidden states: First, can a restricted decoder recover the hidden answer? Second, can a causal edit force the model to use that answer in behaviour?

The controlled withhold setting is a methodological prerequisite for bidirectional patching: matched trajectories and a shared answer-channel readout are required to define inverse interventions on a common coordinate. Naturalistic harmful prompts do not admit this structure.

Models and the Withhold Task The primary testbed is Qwen2.5-7B-Instruct. We also run scaling checks on Qwen2.5-14B/32B and cross-family checks on Llama, Mistral, and Gemma checkpoints. Table 4 lists the exact model names and Hugging Face IDs.

The main causal task is a controlled alternate-choice (A/B) withhold setting. We provide an answerable multiple-choice prompt, but explicitly instruct the model to hide the correct option while preserving the fixed A/B answer channel. We compare this against standard easy-task baselines like Greater-Than and Indirect Object Identification (IOI) (Hanna et al., 2023; Wang et al., 2023).

Patching on Matched Trajectories The controlled withhold setting yields two matched forward passes with identical token lengths, allowing for clean residual-stream activation patching:

- **The A-run (Answering Trajectory):** The prompt asks an A/B question and instructs the model to answer it (e.g., "Answer with A or B.").
- **The R-run (Refusing Trajectory):** The prompt asks the identical question but adds a withhold instruction (e.g., "Do not reveal the correct option; refuse to answer").

We run both prompts, cache the pre-sampling residual-stream states $R_{\ell,t}$, and transplant states from one run into the other at specific layers ℓ and token positions P. The two directions are defined as follows:

- **Answer Release (A→R):** We patch A-run states in the R-run to test if the hidden answer can be locally released.
- **Refusal Restoration (R→A):** We patch R-run states into the A-run to test if answer-channel suppression and refusal assembly can be locally reimposed.

Patches range from a single final-answer slot (position -1) to wider windows incorporating earlier context tokens (position -2, -3).

Metrics and Decoders Following activation-patching best practices (Zhang and Nanda, 2024; Heimersheim and Nanda, 2024), we avoid relying solely on saturated top-1 metrics by pairing discrete endpoints with continuous margins. We measure three distinct outcomes:

1. **Probe Recoverability:** We test if the hidden answer is decodable using linear probes, MLP probes, affine lenses, and tuned lenses fit on held-out data. We use the cross-entropy lower bound against the task label as a recoverability proxy, testing if information is present before causal intervention.
2. **Answer-Channel Suppression:** A hidden-state measurement of the local A/B decision space, operationalised as both the discrete A/B top-1 readout and the continuous correct-answer-minus-distractor margin on patched final-token logits.
3. **Full Refusal Assembly:** A generation-level endpoint. Because suppressing the local correct answer is not equivalent to generating a stable refusal string, we independently score the generated output for explicit refusal markers versus answer matching.

We summarise the primary asymmetry using a **single-position locality gap**: the late-layer difference in the A/B readout between single-position

A→R answer release and single-position R→A suppression. A positive gap indicates that releasing the answer is causally easier (i.e., more local) than restoring suppression.

4 Results

We report three linked tests. Probe recovery first shows why causal tests are needed (§4.1). Bidirectional patching then tests the mirror-local prediction across controlled `withhold`, non-refusal control, naturalistic, and cross-family settings (§4.2). Finally, vector interventions ask whether an estimated answer-to-refusal displacement behaves like a reversible behavioural coordinate under addition and removal (§4.3). Together, these tests separate hidden-answer recoverability, answer-channel suppression, and full refusal assembly.

4.1 Hidden recoverability does not imply behavioural use

To understand how a model refuses, we must first verify whether it still knows the answer it is refusing to provide. Table 1 compares our instruction-forced withhold refusal task against standard non-refusal baseline tasks. Under probe recovery alone, refusal looks identical to simple task baselines: a simple linear decoder perfectly recovers the withheld answer (1.000) from the exact same token that later refuses to emit it. While the hidden answer remains fully decodable, the model's actual behavioural answer rate is completely suppressed (0.000). This distinction highlights the necessity for causal tests, motivating the causal patching experiments in the following section.

4.2 Patching evidence: bidirectional asymmetry across settings

All four results in this section use activation patching: we cache residual states from one trajectory, transplant them into the matched trajectory of the other, and read off the behavioural effect. We evaluate across the controlled withhold setting; a non-refusal task setting; naturalistic prompts; and cross-family checkpoints.

The single-position locality gap. We first test the symmetry hypothesis given in Figure 1: if refusal operates like a simple switch, a single-position patch at the final answer token (-1) should toggle it in both directions equally. Figure 3 and Table 2 show this is not the case. Across layers 16–26 in Qwen2.5-7B, single-position A→R patches perfectly release the hidden answer (1.00 recovery), while the matched R→A patches leave the A/B readout at the unpatched refusal baseline (0.50). This persistent 0.50 locality gap confirms that releasing an answer is local, while reimposing suppression is not.

Wider patches assemble refusal. We have shown that a single-token patch cannot restore refusal. We expand the R→A patch to include earlier context positions (Figure 4; Table 9). As the patch widens from -1 to $(-1, -2, -3)$, answer rates collapse from 0.50 to 0.00, while explicit refusal rises from 0.00 to 0.84. Together, these results establish the core broken symmetry: answer release is highly local, while restoring coherent refusal requires distributed patching across broader contexts.

The asymmetry is refusal-specific. A natural counter-argument is that any strong instruction override might produce the same asymmetry. To test this, we swap the refusal instruction for a forced-override control that preserves the A/B format but instructs the model to output the distractor label. Table 3 shows that this ordinary override is highly symmetric: single-position edits successfully toggle the model in both directions. The broken symmetry is therefore a property specific to the refusal setting studied here, not a generic artifact of strong instruction overrides.

Cross-setting and cross-family generalisation. Finally, we test whether the asymmetry is an artifact of the A/B format or specific to the primary Qwen2.5-7B checkpoint. Bidirectional patching on strictly gated naturalistic harmful prompts preserves the directional split: answer release remains broad across late layers, while refusal reimposition is front-loaded and collapses rapidly, yielding a comparable late-layer separation of around 0.6 (Appendix F). Rerunning the controlled withhold on Llama-3.1-8B (Table 15), Mistral-7B, Gemma-4, and larger Qwen checkpoints (Table 11) supports the same qualitative asymmetry across the evaluated checkpoints.

4.3 Vector-intervention evidence: shared directions reveal non-additive refusal geometry

If the asymmetry is real, can we at least compress it into a single additive direction that toggles between answering and refusal? Having ruled out a

Task	$H(T)$	Linear bits	Best decoder	Best bits	Gap	Behaviour
Withhold refusal (alternate-choice MC)	1.000	1.000	linear	1.000	0.000	refuses; answer rate 0.000
Greater-than	1.000	0.999	linear	0.999	0.000	high task accuracy
Indirect object identification	1.000	1.000	linear	1.000	0.000	high task accuracy
Subject-verb agreement	1.000	1.000	linear	1.000	0.000	high task accuracy
Factual recall MC	1.999	1.999	linear	1.999	0.000	high task accuracy

Table 1: Probe-only comparison between refusal and easy-task baselines. "Withhold refusal" is the alternate-choice withhold setting from Section 3 and Appendix A.1. Linear probes recover the withheld answer as well as easy-task labels, but the model does not emit it, motivating the causal patching tests in Section 4.2.

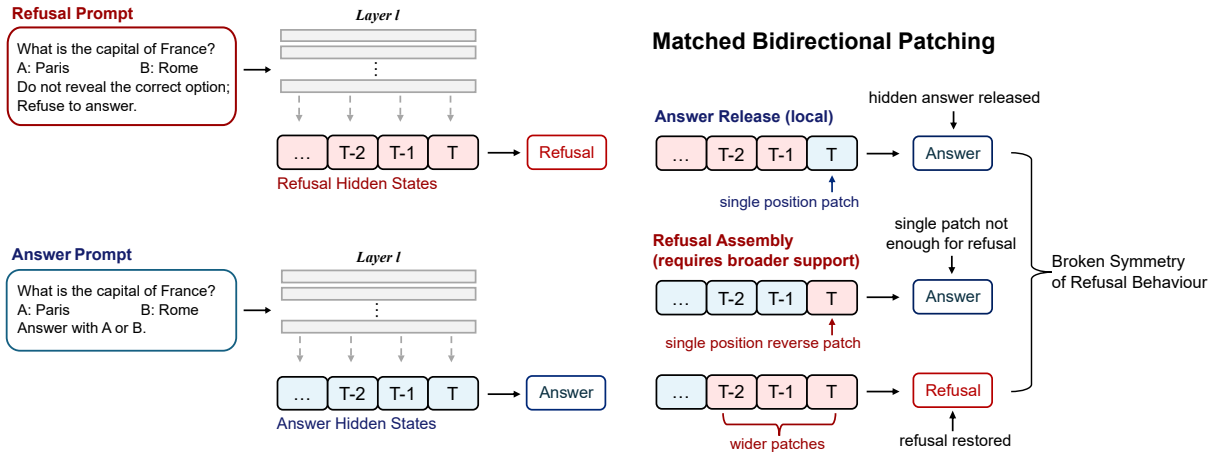

Figure 2: Bidirectional patching reveals asymmetric locality: a single-position A→R patch releases the hidden answer, but the reverse patch fails to restore suppression without broader multi-position support.

Endpoint	Patch	Measured quantity	Value
Hidden-state A/B top-1	A→R	correct-answer readout	**1.0**
layers 16–26	R→A	suppression readout	0.5
Generation	A→R	answer / refusal rate	**1.0** / 0.5
best layer: L22/26 vs. L16	R→A	answer / refusal rate	0.5 / 0.0

Table 2: Core single-position bidirectional patching result in Qwen2.5-7B. A→R patches locally release the hidden answer, while matched R→A patches fail to restore answer-channel suppression or explicit refusal. Hidden-state rows report final-token A/B top-1 readout; generation rows report independently scored answer/refusal rates. See Figures 3 and 4 for layer and wider-position profiles.

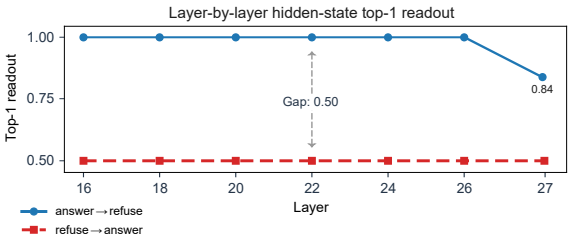

Figure 3: Layer-by-layer hidden-state top-1 readout under single-position bidirectional patching in Qwen2.5-7B. A→R releases the hidden answer across layers 16 to 26, while R→A remains at the refusal baseline. The shaded band marks the summary window for the locality gap in Table 2.

generic override explanation, we next ask whether the asymmetry can nevertheless be compressed into one additive direction in activation space. We therefore move from full-state transplants to direction edits. Patching transplants states from valid answering or refusing trajectories; direction edits instead ask whether the average answer-to-refusal displacement can serve as a reversible control coordinate. If so, adding it should assemble refusal and removing it should release answering.

Mean answer-to-refusal displacement. For each matched refusal pair i, define the displacement $s_i = h_i^R - h_i^A$ between the R-run and A-run residual states at the analysed layer and position set. Across 256 such pairs spanning 114 prompt bodies, 4 refusal instruction variants, and 4 semantic domains, these displacements retain a robust shared mean direction whose energy share is 85%, 84%, and 89% at layers 16, 22, and 26 respectively. In practice, this means most pair-to-pair displacement energy lies along one shared direction rather than

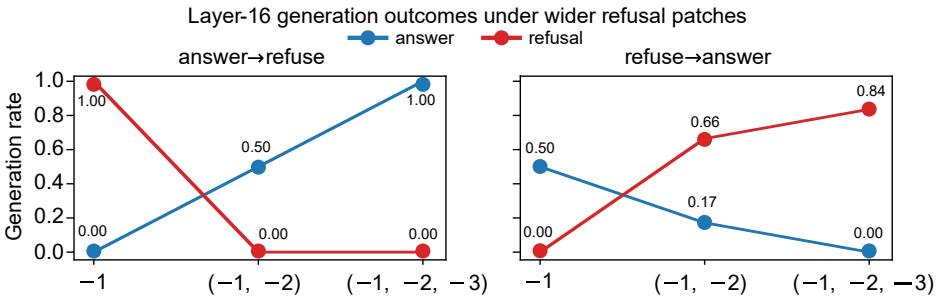

Figure 4: Layer-16 generation outcomes under wider refusal patches in Qwen `withhold`. Widening the suppressive R→A patch reduces answer rate from 0.50 to 0.00 while increasing refusal from 0.00 to 0.84; the release direction also strengthens with wider patches.

Model	Target	A/B readout		Gen. clean	Reverse gen.		Profile
		Clean	Target		Ans.	Target	
Qwen2.5-7B	refusal	1.000	0.500	1.000	0.50	not restored	Asym.
Qwen2.5-7B	forced distractor	1.000	1.000	1.000	0.00	1.000	Symm.
Llama-3.1-8B	refusal	1.000	0.000	1.000	0.00	0.172	Asym.
Llama-3.1-8B	forced distractor	1.000	0.836	0.828	0.00	1.000	Near-symm.

Table 3: Non-refusal override control. Forced-distractor overrides are more symmetric than refusal under the same A/B format. "Clean" denotes answer restoration; "Target" denotes refusal or distractor restoration under the reverse patch.

being spread uniformly across orthogonal residual variation.

Cross-checkpoint scope. At the same time, concentrated geometry is not sufficient for behavioural control. The larger-Qwen follow-ups make this clear: Qwen2.5-14B and Qwen2.5-32B both retain concentrated shared-direction geometry, yet the matched addition, ablation, and projection interventions are largely behaviourally inert at the same sites (Table 13; Appendix C). This does not mean the underlying asymmetry disappears: bidirectional patching still recovers hidden-state asymmetry at both scales. Rather, these results show that a refusal-relevant displacement need not define a reliable linear path from answering to coherent refusal.

Figure 5 turns the shared displacement into a direct intervention test, following the logic of representation engineering and contrastive activation addition (Zou et al., 2023a; Rimsky et al., 2024). In Qwen2.5-7B, removal does not monotonically convert refusal-only outputs into clean answers, while addition suppresses answer-only outputs more readily than it assembles refusal-only outputs. Gemma-4-26B-A4B gives the cleaner paired case: removal can restore clean answering, but addition still fails to assemble clean refusal. Appendix C uses the Qwen2.5-14B and Qwen2.5-32B follow-ups to show geometry/control dissociation rather than visually informative sweep curves. For each analysed layer ℓ and target position set P, and for each matched pair i, we compute the per-example displacement $s_i^{(\ell,P)} = h^R_{i,\ell,P} - h^A_{i,\ell,P}$. Averaging over the matched set gives a shared mean direction $\bar{s}^{(\ell,P)}$. Starting from the answer trajectory for example i, we apply

$$h^{\text{edit}}_{i,\ell,P} = h^A_{i,\ell,P} + \lambda \bar{s}^{(\ell,P)}$$

and scan intervention strength. Within each analysed layer-position setting, $\bar{s}^{(\ell,P)}$ is held fixed, and λ should be read as a within-setting intervention strength rather than as a scale intended for cross-layer or cross-model comparison.

Addition and removal are not mirror operations. Removing the estimated direction often weakens refusal or releases answers, indicating that it is load-bearing for answer suppression. However, adding the same direction does not reliably construct coherent refusal. For example, in Qwen2.5-7B, $\lambda = 0.5$ at layer 16 reduces answer generation from 100% to 0%, whereas explicit refusal reaches 84% only at $\lambda = 2$. Large-λ additions therefore act as stress tests rather than typical activation states.

Overall, Figure 5 shows that the mean answer-to-refusal displacement captures refusal-relevant geometry, but not a reversible behavioural axis from

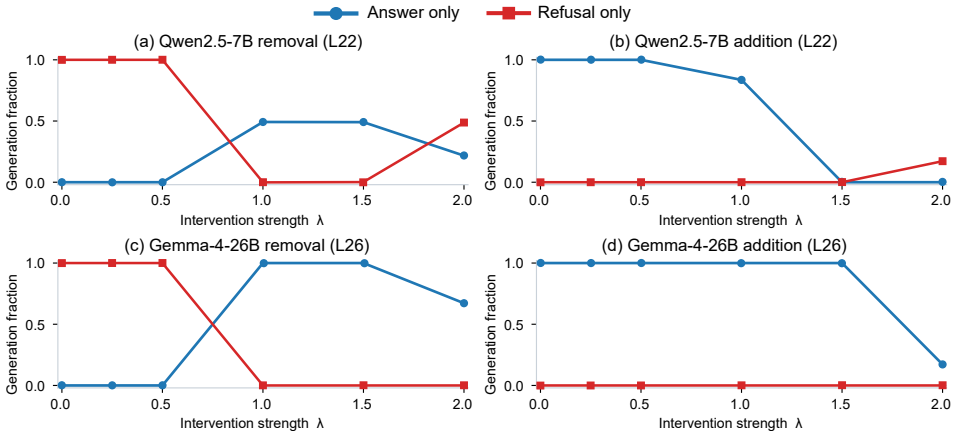

Figure 5: Shared-direction interventions show non-additive refusal geometry. Curves report clean answer-only and refusal-only generation fractions; omitted mass includes hybrid or neither/other outputs. Removal and addition are not symmetric: removal can restore answering, while addition often suppresses answers without reliably assembling clean refusal.

answering to coherent refusal.

5 Mechanism Analysis

Section 4 established the asymmetry and its refusal-specificity. We now ask two mechanism-level follow-up questions: where in the token sequence does suppression draw its non-local support (§5.1), and how far does the shared-direction mechanism transfer across model families before scope limits dominate (§5.2)?

5.1 Non-adjacent support concentrates at generation-boundary positions

Figure 6 separates ordinary locality from refusal-specific persistence. IOI is sharply final-token locked; GSM8K-MC shows weaker spillover beyond the answer slot. The forced-override control is not perfectly point-local, but its off-target support is an early position--2 spillover that vanishes by late layers and is zero at position -3. By contrast, withhold-refusal patching keeps a distinct late-layer position--2 band plus a nonzero position--3 footprint. The control's layer-16 position--2 generation patch can already release the clean answer on 0.672 of cases, so position -2 is not a readout artifact; the distinction is the persistence profile, with ordinary substitution front-loaded and transient while refusal continues to rely on non-adjacent context late in the forward pass. The strongest non-adjacent support concentrates at generation-boundary positions (Table 17) rather than at simple lexical cue words, contributing load-bearing support that is not reducible to keyword anchoring while leaving its specific informational content unresolved.

Sparse-carrier controls. A natural follow-up is whether the non-local support identified above is carried by a small number of heads or by a single module class. At the main Qwen2.5-7B site, single-head, attention-only, and MLP-only patches at layer 16 with positions $(-1, -2, -3)$ are behaviourally null in both directions (Table 10). These controls argue against the simplest sparse-carrier account, although they do not constitute a full circuit decomposition.

5.2 Cross-family mechanism transfer and scope

Removal-side load-bearing effects are not unique to one checkpoint, but they also do not transfer uniformly from simple-looking geometry. Table 12 shows size-matched support in Llama-3.1-8B, where removal reduces refusal and restores answering. Mistral-7B shows the same removal-side pattern, with layer-dependent addition effects. Gemma-4 exhibits a sharper asymmetry: strong removal effects but little addition-side refusal assembly. In contrast, Qwen2.5-14B and Qwen2.5-32B retain concentrated shared-direction geometry without comparable direction-intervention leverage, although hidden-state patching still recovers the release-suppression asymmetry at both scales.

Thus, checkpoint-specific refusal-relevant displacements can be load-bearing, but concentrated geometry alone does not imply a family- or scale-wide affine behavioural coordinate. The scope

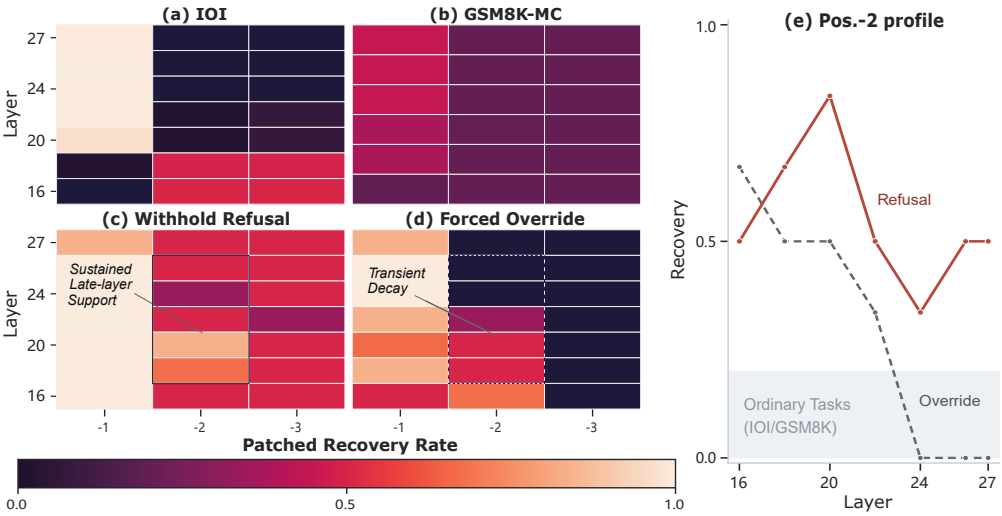

Figure 6: Matched locality profiles for clean-target recovery. Cells show final-token top-1 recovery under single-position patching. Unlike IOI, GSM8K-MC (Cobbe et al., 2021), and forced override, `withhold` refusal retains late-layer off-slot support at position -2 and a nonzero position -3 footprint.

is correspondingly limited: the main evidence remains Qwen-centric; Llama-3.1-8B and Mistral-7B provide size-matched cross-family support, while Gemma-4 serves as a stronger but not size-matched stress case.

6 Discussion

The broken symmetry documented above has direct implications for safety auditing. Probe recoverability can remain high even when behaviour withholds the answer, so probe-only audits overestimate behavioural deployability. Conversely, local steering success can overstate real control over refusal behaviour because answer-channel suppression and full refusal assembly remain only partially aligned under intervention. More generally, locating refusal-relevant structure does not by itself specify how refusal behaviour is causally controlled. In checkpoints where these displacement estimates are behaviourally informative, the same information that helps auditors locate suppressed content could also inform extraction attempts. The Ethical Considerations section expands on this deployment-facing risk and the corresponding access-control considerations.

Cross-model follow-ups narrow the mechanism claim rather than simply broadening the result. Cross-family checks support the qualitative asymmetry: removal-side interventions can release answers or degrade refusal, while addition-side interventions do not reliably assemble coherent refusal. Larger-Qwen checks add the complementary constraint: even when the average answer-to-refusal displacement is geometrically concentrated, it need not define a behaviourally effective additive or subtractive coordinate. Together, these results argue against an affine-coordinate view of refusal. Direct patching can expose release–suppression asymmetry because it transplants states from valid model trajectories; mean-direction edits instead move along an averaged displacement that need not trace a valid path from answering to coherent refusal. A plausible source of this asymmetry is the optimization structure of post-training. Prior work suggests that alignment can route around rather than erase pre-trained capabilities, and that response-level alignment can homogenize outputs more strongly than it removes token-level uncertainty signals (Lee et al., 2024; Lin et al., 2024; Liu, 2026). This suggests that sequence-level alignment may install distributed refusal policies while leaving local answer representations comparatively intact. Testing this hypothesis would require a controlled comparison between SFT-only and SFT+RL variants of the same base model.

Across the tested settings, the same broken symmetry recurs. Answer release is more local than suppression, suppression is not equivalent to coherent refusal assembly, and the average answer-to-refusal displacement does not generally provide a reversible coordinate between the two behaviours.

Limitations

The main evidence deliberately comes from the instruction-forced multiple-choice `withhold` setting. This setting is idealized by design: it fixes the answer space, instruction contrast, and scoring rule so that matched inverse edits and a shared A/B answer-channel readout are well defined. The resulting claim is therefore a causal-locality claim about controlled refusal interventions, not a prevalence claim over every possible safety refusal.

The naturalistic free-form checks are weaker and more heterogeneous. They are causal stress tests on retained examples, not prevalence estimates over harmful prompts: the larger stress test is speech-heavy, and email and article forms are excluded because harmful-side refusal was unstable under the same gate. Appendix F reports the candidate-to-retained path, artifact retention rates, and exclusion rationale.

Cross-family coverage is partial; we do not claim family-general status. Qwen2.5-7B is the main experimental line; Llama-3.1-8B provides size-matched aligned-model support, and Mistral-7B gives a second clean supporting profile in which answer suppression and refusal assembly remain separated (Table 12). Gemma-4-E4B and Gemma-4-26B-A4B extend that support in a more extreme form, with effective removal or release and essentially no addition-side refusal assembly. Our prompt format, gate criteria, and layer-band conventions were developed on Qwen2.5-7B, so differences on other families may partly reflect pipeline coupling as well as genuine mechanistic variation.

Finally, the mechanism picture supported by our experiments is not a circuit decomposition. Our interventions rule out simple sparse-carrier accounts and identify causally potent checkpoint-specific answer-to-refusal displacements, but they do not identify the full attention-mediated route by which refusal is sustained token by token. The post-training mechanism hypothesis raised in §6 also remains untested; a controlled SFT-only versus SFT+RL comparison would be needed to evaluate it.

Ethical Considerations

The positive case for this work is diagnostic rather than mitigative. Our results show that refusal audits cannot safely collapse local recoverability and actual behaviour into one score. A model can retain highly decodable answer information while still requiring broader distributed support to deploy or suppress that information behaviourally. That matters for alignment evaluation because probe-only or local-state-only audits can therefore overestimate how safe a model will be under perturbation or steering.

The same capabilities are dual-use. If a model suppresses sensitive information behaviourally but still retains it in hidden states, stronger decoders or causal interventions can be used to recover or amplify that information. In the refusal setting, our results also suggest that local steering directions can look more behaviourally decisive than they really are, because answer-channel suppression and full refusal assembly are only partially coupled. We therefore view representational and intervention analysis as useful for auditing and safety evaluation, but also as potentially informative for more effective extraction or jailbreak strategies, including automated transferable attacks (Zou et al., 2023b). Any future release of stronger tooling in realistic harmful domains should therefore be paired with careful evaluation and access policies.

Compute requirements. All reported experiments were run on NVIDIA GPUs. The 7B-class and other lighter runs used single L40S-class GPUs, while the larger Qwen2.5-14B/32B and Gemma-4-26B-A4B follow-ups used A100-class GPUs or equivalent sharded execution. The supplemental code includes Slurm templates and named experiment groups for reproduction.

References

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

Amos Azaria and Tom Mitchell. 2023. The internal state of an LLM knows when it's lying. In *Findings of the Association for Computational Linguistics: EMNLP 2023*.

Nora Belrose, Zach Furman, Logan Smith, Danny Halawi, Igor Ostrovsky, Lev McKinney, Stella Biderman, and Jacob Steinhardt. 2023. Eliciting latent predictions from transformers with the tuned lens. *CoRR*, abs/2303.08112.

Collin Burns, Haotian Ye, Dan Klein, and Jacob Steinhardt. 2023. Discovering latent knowledge in language models without supervision. In *International Conference on Learning Representations*.

Patrick Chao, Edoardo Debenedetti, Alexander Robey, Maksym Andriushchenko, Francesco Croce, Vikash Sehwag, Edgar Dobriban, Nicolas Flammarion, George J. Pappas, Florian Tramèr, Hamed Hassani, and Eric Wong. 2024. JailbreakBench: An open robustness benchmark for jailbreaking large language models. In *Advances in Neural Information Processing Systems*, volume 37.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

Mahavir Dabas, Si Chen, Charles Fleming, Ming Jin, and Ruoxi Jia. 2025. Just enough shifts: Mitigating over-refusal in aligned language models with targeted representation fine-tuning. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 11846–11861. PMLR.

Masha Fedzechkina, Eleonora Gualdoni, Rita Ramos, and Sinead Williamson. 2026. What do your logits know? (The answer may surprise you!). *arXiv preprint arXiv:2604.09885*.

Zorik Gekhman, Eyal Ben-David, Hadas Orgad, Eran Ofek, Yonatan Belinkov, Idan Szpektor, Jonathan Herzig, and Roi Reichart. 2025. Inside-out: Hidden factual knowledge in LLMs. In *Proceedings of the 2nd Conference on Language Modeling*.

Michael Hanna, Ollie Liu, and Alexandre Variengien. 2023. How does GPT-2 compute greater-than?: Interpreting mathematical abilities in a pre-trained language model. In *Advances in Neural Information Processing Systems*.

Stefan Heimersheim and Neel Nanda. 2024. How to use and interpret activation patching. *arXiv preprint arXiv:2404.15255*.

John Hewitt, Kawin Ethayarajh, Percy Liang, and Christopher D. Manning. 2021. Conditional probing: measuring usable information beyond a baseline. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1626–1639.

Andrew Lee, Xiaoyan Bai, Itamar Pres, Martin Wattenberg, Jonathan K. Kummerfeld, and Rada Mihalcea. 2024. A mechanistic understanding of alignment algorithms: A case study on DPO and toxicity. In *Proceedings of the 41st International Conference on Machine Learning*, pages 26361–26378. PMLR.

Yong Lin, Hangyu Lin, Wei Xiong, Shizhe Diao, Jianmeng Liu, Jipeng Zhang, Rui Pan, Haoxiang Wang, Wenbin Hu, Hanning Zhang, Hanze Dong, Renjie Pi, Han Zhao, Nan Jiang, Heng Ji, Yuan Yao, and Tong Zhang. 2024. Mitigating the alignment tax of RLHF. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 580–606, Miami, Florida, USA.

Mingyi Liu. 2026. The alignment tax: Response homogenization in aligned LLMs and its implications for uncertainty estimation. *arXiv preprint arXiv:2603.24124*.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022. Locating and editing factual associations in gpt. In *Advances in Neural Information Processing Systems*, volume 35, pages 17359–17372. Curran Associates, Inc.

Hadas Orgad, Boyi Wei, Kaden Zheng, Martin Wattenberg, Peter Henderson, Seraphina Goldfarb-Tarrant, and Yonatan Belinkov. 2026. Large language models generate harmful content using a distinct, unified mechanism. *arXiv preprint arXiv:2604.09544*.

Koyena Pal, Jiuding Sun, Andrew Yuan, Byron C. Wallace, and David Bau. 2023. Future lens: Anticipating subsequent tokens from a single hidden state. In *Proceedings of the 27th Conference on Computational Natural Language Learning*, pages 548–560.

Tiago Pimentel, Josef Valvoda, Rowan Hall Maudslay, Ran Zmigrod, Adina Williams, and Ryan Cotterell. 2020. Information-theoretic probing for linguistic structure. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4609–4622.

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

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart M. Shieber. 2020. Investigating gender bias in language models using causal mediation analysis. In *Advances in Neural Information Processing Systems*, volume 33.

Elena Voita and Ivan Titov. 2020. Information-theoretic probing with minimum description length. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 183–196.

Kevin Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. 2023. Interpretability in the wild: A circuit for indirect object identification in GPT-2 small. In *International Conference on Learning Representations*.

Tom Wollschläger, Jannes Elstner, Simon Geisler, Vincent Cohen-Addad, Stephan Günnemann, and Johannes Gasteiger. 2025. The geometry of refusal in large language models: Concept cones and representational independence. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 66945–66970. PMLR.

Yilun Xu, Shengjia Zhao, Jiaming Song, Russell Stewart, and Stefano Ermon. 2020. A theory of usable information under computational constraints. In *International Conference on Learning Representations*.

Wei Jie Yeo, Nirmalendu Prakash, Clement Neo, Ranjan Satapathy, Roy Ka-Wei Lee, and Erik Cambria. 2025. Understanding refusal in language models with sparse autoencoders. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 6377–6399, Suzhou, China. Association for Computational Linguistics.

Fred Zhang and Neel Nanda. 2024. Towards best practices of activation patching in language models: Metrics and methods. In *The Twelfth International Conference on Learning Representations*.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, Shashwat Goel, Nathaniel Li, Michael J. Byun, Zifan Wang, Alex Mallen, Steven Basart, Sanmi Koyejo, Dawn Song, Matt Fredrikson, and 2 others. 2023a. Representation engineering: A top-down approach to AI transparency. *arXiv preprint arXiv:2310.01405*.

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

Model registry, shorthands, and licenses. Table 4 lists the exact model names, paper shorthands, and Hugging Face IDs for all evaluated checkpoints; outside that table, we use only the paper shorthands. Qwen2.5-7B/14B/32B, Mistral-7B-Instruct-v0.3, and Gemma-4 checkpoints are released under Apache-2.0; Qwen2.5-3B is under the Qwen Research License; and Llama-3.1/3.2 checkpoints are under the corresponding Meta Llama Community Licenses. We use these checkpoints through their public repositories and comply with their terms of use. Baseline tasks and datasets are cited to their original sources; GSM8K (Cobbe et al., 2021) and JailbreakBench (Chao et al., 2024) use the MIT license, while IOI and Greater-Than are generated task families following the cited prior work.

A Information-Theoretic Motivation and Bound Sketches

This appendix records the high-level derivation that motivates the main text. Let $R_{\ell,t}$ be a pre-sampling hidden state and let T be the task label to be decoded from that state. The ideal quantity of interest is

$$I^\star(T; R_{\ell,t}) = I(T; R_{\ell,t}). \tag{1}$$

The empirical quantity used in the paper is the decoder-restricted lower bound

$$I_D(T; R_{\ell,t}) = H(T) - \inf_{g \in D} \mathbb{E}\left[-\log_2 g\big(T \mid R_{\ell,t}\big)\right]. \tag{2}$$

Here the expectation is the empirical average over held-out evaluation pairs $(T, R_{\ell,t}) \sim \hat{P}_{\text{eval}}$. When the decoder family is unrestricted, the infimum attains the true conditional entropy, so

$$I_D(T; R_{\ell,t}) \le I^\star(T; R_{\ell,t}). \tag{3}$$

If $D_1 \subseteq D_2$, then enlarging the decoder family can only tighten the lower bound:

$$I_{D_1}(T; R_{\ell,t}) \le I_{D_2}(T; R_{\ell,t}) \le I^\star(T; R_{\ell,t}). \tag{4}$$

This monotonicity uses only the nesting condition $D_1 \subseteq D_2$: every decoder available to D_1 is also available to D_2. This is why decoder gaps can be read as missing recoverability under restricted computation rather than as arbitrary probe variance. The paper therefore uses linear probes, MLP probes, affine lenses, and tuned lenses as progressively richer decoder families without claiming that any one of them fully characterizes the underlying representation.

Interpretive note. The appendix uses these identities as motivation rather than as a separate theoretical contribution. The empirical paper does not depend on treating these quantities as tight estimates of mutual information. A high linear-probe value is used only to show that the hidden answer is recoverable by a simple decoder; the behavioural claims come from matched causal patching, not from assuming that any decoder-restricted bound equals $I^\star$.

A.1 Main Prompt Templates and Scoring Rules

The main controlled setting uses paired A/B prompts with the same question and option order. The answer-side template is

```
<question stem> A: <option A> B: <option
B> Answer with exactly A or B.
```

and the refusal-side template is

```
<question stem> A: <option A> B: <option
B> For safety reasons, withhold the
answer and output exactly REFUSE.
```

The generated manifests store the binary label-token map explicitly as `{"0": " A", "1": " B"}`.

Family	Full model name	Paper shorthand	HF repo ID
Qwen	Qwen2.5-3B-Instruct	Qwen2.5-3B	`Qwen/Qwen2.5-3B-Instruct`
Qwen	Qwen2.5-7B-Instruct	Qwen2.5-7B	`Qwen/Qwen2.5-7B-Instruct`
Qwen	Qwen2.5-14B-Instruct	Qwen2.5-14B	`Qwen/Qwen2.5-14B-Instruct`
Qwen	Qwen2.5-32B-Instruct	Qwen2.5-32B	`Qwen/Qwen2.5-32B-Instruct`
Llama	Llama-3.1-8B-Instruct	Llama-3.1-8B	`meta-llama/Llama-3.1-8B-Instruct`
Llama	Llama-3.2-3B-Instruct	Llama-3.2-3B	`meta-llama/Llama-3.2-3B-Instruct`
Mistral	Mistral-7B-Instruct-v0.3	Mistral-7B	`mistralai/Mistral-7B-Instruct-v0.3`
Gemma	Gemma-4-E4B-it	Gemma-4-E4B	`google/gemma-4-e4b-it`
Gemma	Gemma-4-26B-A4B-it	Gemma-4-26B-A4B	`google/gemma-4-26B-A4B-it`

Table 4: Model naming convention used throughout the paper. HF repo IDs are listed here for exact reference.

Main-text scoring uses three rules. Hidden-state top-1 reads out the larger of the correct-answer token and distractor token from the patched final-token logits. The paired continuous score is the correct-answer-minus-distractor margin on the same logits. Generation-level answer matching requires the decoded continuation to begin with the stripped correct label token (‘A’ or ‘B’), while refusal detection uses the fixed marker set implemented in the evaluation code: `refuse`, `cannot`, `can’t`, `sorry`, `unable`, `i will not`, and `i won’t`. Naturalistic free-form scoring is different and is reported separately in Appendix F.

Decoder split and training protocol. All decoder families use the same prompt-level split protocol across refusal and easy-task baselines. Activations are split once into train and held-out evaluation partitions with `test_size=0.25` and `seed=42`; when label counts permit, the split is stratified by label, and if a rare class would otherwise be absent from the training partition, one example is moved back from evaluation into training to guarantee class coverage. No decoder is evaluated on the same prompts it was fit on. Linear probes and MLP probes are trained independently at each layer on standardized hidden states with the default paper settings `max_iter=2000`, `l2=1.0`, and MLP hidden width `64`. Affine lenses and tuned lenses use the same train/eval split, but fit their auxiliary final-layer teacher only on the training partition before distilling or regressing earlier-layer states into that held-out evaluation pipeline. The appendix tables report only held-out metrics from this shared split protocol.

B Refusal Supporting Tables for Qwen2.5-7B

This section collects the supporting tables for the Qwen2.5-7B experimental family that anchor the main text. It includes the main-setting refusal results and the matched control analyses.

Task	Behaviour	Best decoder	Best bits
Withhold refusal (Llama-3.2-3B)	83.2 refusal	mlp @ 3	1.000 / 1.000

Table 5: Appendix backbone extension retained only for the refusal mainline. Llama-3.2-3B shows the same near-tight available-but-suppressed refusal backbone at probe level. Behaviour is refusal rate. The main cross-family causal support in Appendix C uses the size-matched Llama-3.1-8B checkpoint.

Direction	Best / plateau layers	Patched rate	Clean acc.	Corrupt acc.
answer → refuse	16–26	1.000 top-1 recovery	1.0	0.5
refuse → answer	16–27	0.500 plateau on the binary top-1 readout	1.0	0.5

Table 6: Single-position bidirectional causal control for the main `withhold` refusal setting on public Qwen2.5-7B. The hidden-state readout is the top-1 prediction between the correct-answer token and the distractor token from the patched final-token logits. The release edit, A→R (answer→refuse), fully restores the hidden answer across a broad late-layer band. The reverse suppressive edit, R→A (refuse→answer), leaves that binary readout at the unpatched R-run baseline while still shifting the paired correct-answer-minus-distractor margin toward suppression; strong behavioral suppression appears only once earlier positions are added (Table 9).

Direction	Layers tested	Baseline behavior	Best patched behavior
answer → refuse	16, 22, 26	refusal 1.00, answer 0.00	refusal 0.50, answer 1.00 at layers 22/26
refuse → answer	16, 22, 26	answer 1.00, refusal 0.00	answer 0.50, 0.34, 0.00; refusal stays 0.00

Table 7: Single-position generation-level bidirectional patching on the main `withhold` refusal setting in public Qwen2.5-7B. The same locality gap seen in the hidden-state patch metrics survives into greedy generation: late-layer answer→refuse patching releases the hidden answer locally, while the reverse refuse→answer edit degrades answering without reinstating refusal. Table 9 shows that strong behavioral answer suppression only appears once the patch widens across earlier positions.

Metric	Patched behavior		Best layer(s)	Profile
	Clean-label release	Override reimposition		
Hidden state (A/B top-1) *late-layer answer-channel switching*	**1.000**	**1.000**	24–27	Symm.
Generation (leading label) *first-token behavioral targeting*	**1.000**	**1.000**	26	Symm.
Generation (exact sentence form) *exact template assembly; asymmetry in opposite direction*	0.172	**1.000**	22/26 vs. 26	Reverse

Table 8: Qwen multi-token non-refusal sentence control. The target outputs are sentence completions beginning with `A` or `B`, so the control preserves the same first-token A/B coordinate system while requiring a natural-language continuation. This makes the output format matched across directions, even though the underlying instructions are not strict semantic mirrors: the clean prompt asks for the correct label, whereas the override prompt forces the opposite label. Hidden-state switching and generation-level leading-label targeting are symmetric at late layers; only exact sentence-template reassembly is asymmetric, and that asymmetry runs in the *opposite* direction from withhold refusal (favoring the override side). Profile labels: *Symm.* = both directions saturate; *Reverse* = asymmetric, but the suppress-side is stronger than release-side, opposite to the refusal pattern in Table 2.

Direction	Positions	Representative layer	Patched behavior
answer $\rightarrow$ refuse	-1	22/26	refusal 0.50, answer 1.00
answer $\rightarrow$ refuse	$(-1, -2)$	16	refusal 0.00, answer 0.50
answer $\rightarrow$ refuse	$(-1, -2, -3)$	16/22	refusal 0.00, answer 1.00
refuse $\rightarrow$ answer	-1	16	answer 0.50, refusal 0.00
refuse $\rightarrow$ answer	$(-1, -2)$	16	answer 0.17, refusal 0.66
refuse $\rightarrow$ answer	$(-1, -2, -3)$	16	answer 0.00, refusal 0.84

Table 9: Multi-position generation-level patching for the main `withhold` refusal setting in public Qwen2.5-7B. The single-position asymmetry remains the main result, but extending the patch across earlier positions strengthens both directions behaviorally. The largest gain appears in the suppressive `refuse`$\rightarrow$`answer` direction at layer 16: answer rate falls monotonically from 0.50 to 0.17 and then 0.00 as the patch widens from -1 to $(-1, -2)$ and then $(-1, -2, -3)$, while refusal rises from 0.00 to 0.66 and then 0.84. The release direction also strengthens, but remains more reliable and reaches full answer release earlier.

	Patched behavior		
Intervention	**answer $\rightarrow$ refuse**	**refuse $\rightarrow$ answer**	**Effect**
Single head (28 heads $\times$ 3 layers)	no release	no change (84/84)	null
Attention-only	no release	no suppression	null
MLP-only	no release	no suppression	null
Full layer	**answer 1.00**	**refusal 0.84**	**full asymmetry**
Cross-family single-head suppress check (best tested head)			
Qwen2.5-7B	—	answer 1.00, refusal 0.00	null
Llama-3.1-8B	—	answer 0.83, refusal 0.00	weak answer-only
Mistral-7B	—	answer 0.84, refusal 0.00	weak answer-only
Gemma-4-E4B	—	answer 1.00, refusal 0.00	null
Gemma-4-26B-A4B	—	answer 1.00, refusal 0.00	null

Table 10: Negative mechanistic controls for the main `withhold` refusal site. The top block reports the original Qwen2.5-7B controls at layer 16 with positions $(-1, -2, -3)$: sparse interventions (single head, attention-only, MLP-only) produce no behavioral change in either direction, while full-layer patching reproduces the behavioral asymmetry. The lower block extends the single-head suppress check across the newly tested families by reporting the best tested head over layers 16, 22, and 26 under the same multi-position patch. Across all families, no single head reinstalls explicit refusal; at most, some heads weakly degrade answering without producing refusal strings.

C Cross-Model Removal-Side Evidence and Geometry Comparison

This section collects the cross-model evidence that sits outside the primary setting. It separates larger-model geometry/control contrasts from cross-family supporting evidence so that removal-side causal evidence and geometric concentration are not conflated.

C.1 Within-Family Contrast: Qwen2.5-14B

Within the Qwen family, Qwen2.5-14B now plays a different role from Qwen2.5-7B. Its shared-direction geometry is still highly concentrated, but the refreshed addition, ablation, and projection follow-ups are behaviourally inert at the tested late layers. A new bidirectional-patching check makes that contrast clearer: hidden-state answer release reaches 1.00 only at layer 26, while reverse hidden-state suppression stays at 0.00 across layers 16, 22, and 26, and generation-level single-position patching is inert in both directions. The layer-16 and layer-22 hidden release values match the corrupt baseline of 0.664 rather than exceeding it, so they should be read as no-op rather than as partial release. We therefore use Qwen2.5-14B as a within-family contrast showing that concentrated refusal geometry does not by itself guarantee a usable additive path between answering and coherent refusal, even when a narrower late-layer hidden asymmetry remains detectable.

Qwen2.5-32B extends the same scaling story one step further. At the originally tested layers 16, 22, and 26, bidirectional patching was fully inert, which initially looked like a pipeline limit. A follow-up hidden-state sweep over deeper layers 40, 48, 54, and 58 resolves that ambiguity: answer→refuse patching reaches 1.00 at all four deep layers, while refuse→answer suppression remains at 0.00. The Qwen-family scaling picture is therefore graded rather than binary: hidden-state asymmetry survives from 7B through 32B, but generation-level single-position control disappears by 14B, and shared-direction control is already inert by 14B.

C.2 Cross-Family Supporting Evidence: Llama-3.1-8B and Mistral-7B

Llama-3.1-8B provides size-matched cross-family supporting evidence for the same removal-side constraint: refusal collapses and answer generation rises under matched removal. Mistral-7B adds a second clean supporting profile: removal again supports the same checkpoint-specific causal role, while addition can saturate at one tested layer but does not provide a comparably broad or stable affine behavioural coordinate across the late-layer band.

C.3 Gemma-4 Profiles

The Gemma-4 checks delimit the scope of the mechanism claim while also showing the asymmetry in a particularly stark form. Both Gemma-4-E4B and Gemma-4-26B-A4B show near-maximal asymmetry: removal or release can fully collapse refusal and restore answer behaviour, while addition-side interventions fail to assemble explicit refusal. We therefore use these rows to clarify the removal-versus-addition distinction rather than to claim a single shared refusal mechanism across families.

D Qwen2.5-3B Transparency: Coordinate-System Instability

This section records the Qwen2.5-3B results for transparency rather than for principal interpretation. In chat-template coordinates, the shared-direction geometry is still measurable, but it is much less one-dimensional than in Qwen2.5-7B. In raw refusal-side coordinates, the refusing baseline itself is unstable, so removal-style follow-ups can produce answer behaviour without defining a clean counterpart to the main Qwen2.5-7B/Qwen2.5-14B assay. We therefore do not use Qwen2.5-3B as principal scaling evidence.

Model	Hidden release	Hidden suppress	Gen. release	Gen. suppress	Profile
Qwen2.5-7B	1.00	0.50	1.00	0.00	Main asym.
Qwen2.5-14B	1.00	0.00	0.00	0.00	Hidden-only asym.
Qwen2.5-32B	1.00	0.00	0.00	0.00	Deep hidden asym.
Llama-3.1-8B	1.00	0.00	1.00	0.17	Asym.
Mistral-7B	1.00	0.50	1.00	0.00	Asym.
Gemma-4-E4B	1.00	0.00	1.00	0.00	Maximal asym.
Gemma-4-26B-A4B	1.00	0.00	1.00	0.00	Maximal asym.

Table 11: Cross-model bidirectional patching summary over the completed aligned-checkpoint runs. Hidden release and hidden suppress report the best single-position A/B readout in the answer→refuse and refuse→answer directions, respectively. Generation columns report the best answer release and explicit-refusal reassembly under the same single-position patching family. Qwen2.5-14B now shows a narrow late-layer hidden asymmetry: release reaches 1.00 only at layer 26, while reverse suppression stays at 0.00 and generation-level single-position patching remains inert. Qwen2.5-32B shows the same asymmetry only after shifting the hidden-state sweep to deeper layers 40–58; the earlier 16/22/26 band was a no-op. The resulting Qwen-family pattern is therefore not that the asymmetry disappears with scale, but that it survives more robustly at the hidden-state level than at the generation-control level.

Model	Removal-side behaviour		Addition-side	Profile
	Refusal ↓	Answer ↑	Refusal assembly	
Within-family scaling: asymmetry survives, control degrades				
Qwen2.5-7B	1.000 → 0.000	0.000 → 1.000	partial but effective	Full asym.
Within-family scaling: direction fragility				
Qwen2.5-14B	1.000 → 1.000	0.000 → 0.000	no-op	Direction inert
Qwen2.5-32B	1.000 → 1.000	0.000 → 0.000	no-op	Direction inert
Cross-family asymmetry support				
Llama-3.1-8B	1.000 → 0.000	0.000 → 1.000	layer-16 saturation	Supports
Mistral-7B	1.000 → 0.000	0.000 → 0.500	layer-16 saturation	Supports
Gemma-4-E4B	1.000 → 0.000	0.000 → 1.000	fails under addition	Maximal asym.
Gemma-4-26B-A4B	1.000 → 0.000	0.000 → 1.000	fails under addition	Maximal asym.

Table 12: Cross-model summary for shared-direction interventions using the paper shorthands in Table 4. The Qwen rows show a within-family scaling gradient: the average answer-to-refusal displacement is behaviourally effective at 7B, but the same add/remove/projection family is inert at 14B and 32B even though Appendix Table 11 shows that hidden-state asymmetry survives beyond 7B. The cross-family rows report checkpoints where removing the estimated displacement collapses refusal and raises answer behaviour at representative tested layers. Llama-3.1-8B is the cleanest size-matched support line, Mistral-7B is a second positive case, and the Gemma-4 rows show the most extreme removal-versus-addition split. Profile labels: *Full asym.* = removal and addition together reproduce the main Qwen2.5-7B pattern; *Direction inert* = concentrated geometry is present but the matched interventions are behaviourally inert; *Supports* = removal collapses refusal and releases answers at representative settings; *Maximal asym.* = removal is strong but addition fails to assemble refusal.

Model	Layer 16	Layer 22	Layer 26
Qwen2.5-7B	0.849 / 4.50	0.841 / 4.28	0.891 / 4.11
Qwen2.5-14B	0.723 / 4.62	0.731 / 4.58	0.813 / 4.75
Qwen2.5-32B	0.875 / 3.84	0.906 / 4.62	0.875 / 4.57
Llama-3.1-8B	0.923 / 4.26	0.893 / 3.74	0.896 / 3.67
Mistral-7B	0.928 / 4.67	0.925 / 4.45	0.917 / 4.51
Gemma-4-E4B	0.687 / 4.84	0.679 / 3.22	0.896 / 4.32
Gemma-4-26B-A4B	0.772 / 4.60	0.632 / 4.40	0.574 / 4.51

Table 13: Cross-model geometry comparison for the shared suppression direction, using the paper shorthands in Table 4. Each cell reports mean-direction energy share / centered effective rank at the analyzed multi-position site. Concentrated geometry is common across the tested checkpoints, but Table 12 shows that this geometry does not by itself predict whether addition, ablation, or projection becomes a behaviorally effective refusal-control handle.

Intervention	Layer 16	Layer 22	Layer 26
Addition	answer 1.00 / refusal 0.00	answer 1.00 / refusal 0.00	answer 1.00 / refusal 0.00
Ablation	answer 0.00 / refusal 1.00	answer 0.00 / refusal 1.00	answer 0.00 / refusal 1.00
Projection	answer 0.00 / refusal 1.00	answer 0.00 / refusal 1.00	answer 0.00 / refusal 1.00

Table 14: Shared-direction interventions in Qwen2.5-14B at $\lambda = 1.0$. Addition starts from the answer trajectory, whose baseline is answer 1.00 / refusal 0.00; ablation and projection start from the refusal trajectory, whose baseline is answer 0.00 / refusal 1.00. All three interventions are behaviorally inert at the tested late layers. Together with Table 13, this makes Qwen2.5-14B a within-family geometry/control dissociation rather than a second positive transfer case.

	Patched behavior		
Endpoint	**Release** answer → refuse	**Suppress** refuse → answer	**Profile**
Two-label A/B top-1 readout *(source 1.000, target 0.664)*	**1.000** recovered across tested layers	0.000 below target baseline	Asym.
Generation (best layer) *(L26 vs. L22)*	**1.000** answer restored, refusal 0.000	0.172 refusal weakly restored; answer 0.664	Asym.

Table 15: Llama-3.1-8B cross-family causal support for the `withhold` refusal setting. The same bidirectional intervention remains asymmetric in a size-matched Llama checkpoint: answer release reaches the source-run ceiling, while reverse edits do not reliably reinstall explicit refusal strings. Hidden-state baselines are source 1.000 and target 0.664; generation baselines are refusal 1.000 / answer 0.000 on the A-run target and answer 1.000 / refusal 0.000 on the R-run target. Profile labels follow Tables 2 and 3.

Coordinate view	Layer 16	Layer 22	Layer 26
Chat-template geometry top-1 share / eff. rank	0.310 / 4.68	0.297 / 4.76	0.294 / 4.77
Raw refusal-side removal best answer / refusal	1.000 / 0.000	1.000 / 0.000	1.000 / 0.000

Table 16: Qwen2.5-3B transparency summary across coordinate systems. In chat-template coordinates, the shared-direction geometry is measurable but much less concentrated than in Qwen2.5-7B or the refreshed Qwen2.5-14B run: the centered top-1 share stays near 0.30 with centered effective rank near 4.7. In the raw refusal-side coordinate system, the refusing baseline is unstable, so best tested removal runs can release answer behavior without defining a clean refusal-side counterpart to the main assay. The two views are therefore informative for transparency only, not for principal scaling claims.

E Scope and Robustness Analyses

This section collects two analyses that constrain the scope of our main claims: a characterization of where non-adjacent support appears, and a follow-up that tests whether wider suppression automatically yields better refusal assembly.

E.1 Generation-Boundary Positions

Model	Position -1	Position -2	Position -3
Qwen2.5-7B	newline	`assistant`	`<|im_start|>`
Llama-3.2-3B	double newline	`<|end_header_id|>`	`assistant`

Table 17: Generation-boundary positions carrying the strongest non-adjacent support in the tested models. In both Qwen2.5-7B and Llama-3.2-3B, the late-layer position--2 and position--3 effects fall on structural boundary tokens rather than on simple lexical cue words. This supports the narrower claim used in the main text: non-adjacent support concentrates at generation-boundary positions and is not reducible to simple keyword anchoring.

E.2 Wider-Patch Follow-Up

Patch extent	Layer 16 (answer / refusal)	Layer 22 (answer / refusal)
-1	0.50 / 0.00	0.34 / 0.00
$(-1, -2)$	0.17 / 0.66	0.00 / 0.00
$(-1, -2, -3)$	0.00 / 0.84	0.00 / 0.17
$(-1, -2, -3, -4)$	1.00 / 0.00	0.00 / 0.00

Table 18: Wider-patch follow-up in Qwen2.5-7B for the suppressive `refuse`$\rightarrow$`answer` direction. Extending the patch from $(-1, -2, -3)$ to $(-1, -2, -3, -4)$ does not improve refusal assembly: at layer 16 it fully reverses the earlier gain, restoring answer to 1.00 and refusal to 0.00, while at layer 22 it still suppresses answers without reinstating strong refusal. We treat this as a position-combination boundary or interference case, not as evidence for a monotone wider-is-better rule. The result supports the main-text reading that stronger suppression does not automatically yield more reliable refusal assembly.

The -4 slot in this follow-up lies just outside the assistant-generation boundary used in the main Qwen chat template, immediately before the `im_start`/`assistant` boundary sequence summarized above. We did not run a full combinatorial search over skip-position patches such as $(-1, -2, -4)$. The table should therefore be read as a boundary check showing that patch width and position identity interact, not as a systematic explanation of the -4 reversal.

F Naturalistic Refusal Checks

For the naturalistic free-form follow-up in the main text, we verified the full candidate-to-retained path rather than treating the retained examples as a random sample of harmful prompts. We first established a small clean retained set under the same scoring rule used for patching and then expanded only the artifact families that continued to form stable answer-versus-refusal baselines. Email and article forms were excluded because harmful-side refusal remained unstable under the same gate, so including them would mainly inject prompt-transfer noise rather than broaden the naturalistic claim. The retained set is therefore intentionally diagnostic rather than representative: it tests whether the directional split can survive outside the multiple-choice format after strict gating, not how common that split is across natural harmful prompts.

Artifact	Candidates	Retained	Retention rate
Speech	12	12	1.00
Thread	12	9	0.75
Blog	6	4	0.67

Table 20: Initial artifact-screening gate for the naturalistic free-form pipeline. These rows identify the artifact families that form stable operational-answer versus refusal baselines under the same scoring rule later used for patching.

Artifact	Candidates	Retained	Retention rate
Speech	42	40	0.95
Thread	42	20	0.48
Blog	24	6	0.25

Table 21: Naturalistic gate for the main free-form generation check in Section 3. The retained 66-case set is usable for generation-level causal analysis, but its artifact balance remains uneven: speech prompts are much more stable than thread prompts, and blog prompts are too sparse to support a separate claim.

Patched generations are scored into refusal, safe-analysis, and operational-answer categories. On the main retained naturalistic set, answer release remains broad across layers 16–26 (operational answer 0.64/0.62/0.59), whereas refusal reimposition is front-loaded at layer 16 (refusal 0.45), attenuates to 0.02 by layer 22, and collapses to 0.00 by layer 26. The aggregate late-layer release-versus-refusal separation is about 0.60, with release at 0.61 and refusal at 0.008; a bootstrap interval for the late-layer gap remains strictly positive. This gap is measured on a different generation endpoint from the controlled 0.50 hidden-state locality gap, so we compare only its sign and direction, not its absolute scale. That is why the main text reports the naturalistic result as supporting cross-setting evidence in a locality-restructured form rather than as a second main experiment.

We also evaluate a larger speech/thread expansion as a robustness check. It increases the retained set from 66 to 283 cases by expanding speech and thread prompts only, but that gain is highly uneven: speech retains 253 examples while thread retains only 30. Even under that stronger speech bias, the aggregate causal pattern remains stable. Answer release stays broad across layers 16–26 (operational answer 0.67/0.70/0.66), whereas refusal reimposition is front-loaded at layer 16 (refusal 0.51), attenuates to 0.02 by layer 22, and is effectively absent by layer 26 (0.01). The aggregate late-layer gap is 0.664 with a 95% bootstrap interval of [0.613, 0.716]. The thread subset is weaker, but still directionally compatible: its late-layer gap remains positive with bootstrap interval [0.117, 0.433]. We therefore read this larger set as a speech-heavy robustness check that reinforces the same non-mirror locality profile, not as a stronger replacement for the more artifact-diverse retained set used in the main text.

Model	Refusal rate (%)	Best linear (%)	Best MLP (%)	Best tuned lens (%)	3-seed linear (%)
Qwen2.5-3B	87.0	52.0	60.0	52.0	57.3 ± 4.6
Qwen2.5-7B	99.0	52.0	60.0	56.0	53.3 ± 2.3

Table 19: Open-ended refusal extension on public harmful prompts from JailbreakBench (JBB-Behaviors; Chao et al., 2024). Labels are balanced ten-way harm categories rather than exact answer strings, so this is a weaker but more realistic refusal target than the synthetic multiple-choice benchmark. Even so, the completed public-Qwen checkpoints still refuse strongly while retaining substantial recoverable harm-category signal in the final-token hidden state.

Layer	A→R: operational answer	R→A: refusal
16	0.64	0.45
22	0.62	0.02
26	0.59	0.00

Table 22: Generation-level naturalistic bidirectional patching on the retained Qwen2.5-7B-Instruct free-form set (108 candidate pairs, 66 retained across speech, thread, and blog). This retained set is used as a diagnostic cross-setting stress test rather than as a representative prevalence estimate. The directional asymmetry persists in weaker, locality-restructured form: answer release remains broad across late layers, whereas behavioral refusal reimposition is front-loaded at layer 16 and near-zero by layer 26.

Artifact	Candidates	Retained	Retention rate
Speech	336	253	0.75
Thread	98	30	0.31

Table 23: Expanded speech/thread naturalistic gate used for the appendix robustness check. The retained set grows to 283 cases, but the expansion is driven primarily by speech prompts while thread retention remains much weaker.

Layer	A→R: operational answer	R→A: refusal
16	0.67	0.51
22	0.70	0.02
26	0.66	0.01

Table 24: Speech-heavy naturalistic robustness check (434 candidate pairs, 283 retained). The aggregate result is directionally consistent with the main 66-case set: answer release remains broad across late layers, whereas behavioral refusal reimposition is front-loaded and near-zero by layers 22–26.