Where Does LoRA Write Counterfactual Reasoning in Large Language Models?

Anonymous ACL submission

Abstract

Large language models (LLMs) struggle with counterfactual reasoning, achieving near-chance accuracy on established benchmarks. Low-Rank Adaptation (LoRA) fine-tuning improves performance to 89–95% on CLadder, but the mechanistic basis remains unclear. We investigate *where* LoRA writes counterfactual reasoning across **five LoRA-finetuned models spanning three scales (7B–72B) and three architecture families** (Qwen2.5-7B/14B/72B, Llama3-8B, DeepSeek-R1-Distill-Qwen-7B), using three complementary signals—representation similarity, linear probing, and bidirectional activation patching—that co-localize to the same critical layer band in every model. Two further analyses go beyond confirming that upper layers carry semantics: (i) a **token-shuffle control** preserves activation statistics while destroying token alignment and completely abolishes patching recovery (97%→53% on Qwen-7B; full collapse in all five models), establishing that LoRA writes *structured, position-specific* information rather than generic distributional shifts—a general-purpose diagnostic for any patching study; (ii) **component-level decomposition** into Attention-only and MLP-only patching reveals that layer-level localization is consistent across architectures, but the component composition is architecture-specific—Qwen is Attention-leaning in the critical band (with the Attention-dominant share rising with scale: 56%/61%/66% at 7B/14B/72B), DeepSeek-R1 is MLP-dominant, and Llama is balanced. The same input–output mapping can therefore be installed via different component pathways, complicating "one mechanism fits all" accounts of how PEFT writes reasoning capabilities.

1 Introduction

Counterfactual reasoning—reasoning about hypothetical alternatives to observed events ("Would the patient have recovered if treated differently?")—is fundamental to causal understanding, planning, and decision-making (Pearl, 2009).

Despite progress in natural language understanding, LLMs exhibit striking deficiencies on this task. On CLadder (Jin et al., 2023), which evaluates counterfactual inference under structural causal models, even GPT-4 reaches only ∼65% accuracy and 7–8B-parameter models perform near chance (∼50%). The pattern persists across scales and is consistent with broader multi-step reasoning evaluations: models that improve substantially on arithmetic and compositional suites still fail on intervention-style causal questions (Cobbe et al., 2021; Srivastava et al., 2023; Suzgun et al., 2023; Wei et al., 2022), indicating a bottleneck beyond generic multi-step inference.

Low-Rank Adaptation (LoRA) (Hu et al., 2022) offers a parameter-efficient remedy: it raises counterfactual accuracy from ∼50% to 89–95% on CLadder (up to 99.7% on CounterBench) in every model we test, including Qwen-72B. This raises a mechanistic question: *what does LoRA change, and where in the model does this change occur?* Localizing where capabilities are written enables targeted editing, cross-model capability transfer, and interpretability of which architectural regions support which cognitive functions (Meng et al., 2022; Geva et al., 2023).

Figure 1 summarizes our pipeline. We use three complementary methods: **(1) Representation Similarity** for where representations diverge; **(2) Linear Probing** for where decision signals become decodable; and **(3) Activation Patching** to establish causal relationships (Zhang and Nanda, 2024; Heimersheim and Nanda, 2024). Crucially, we add a **token-shuffle control** that rules out activation-statistic confounds, and a **component-level decomposition** into Attention- and MLP-only patching. We validate across five models at three scales (7B/14B/72B); for Qwen2.5-72B we apply a cache-

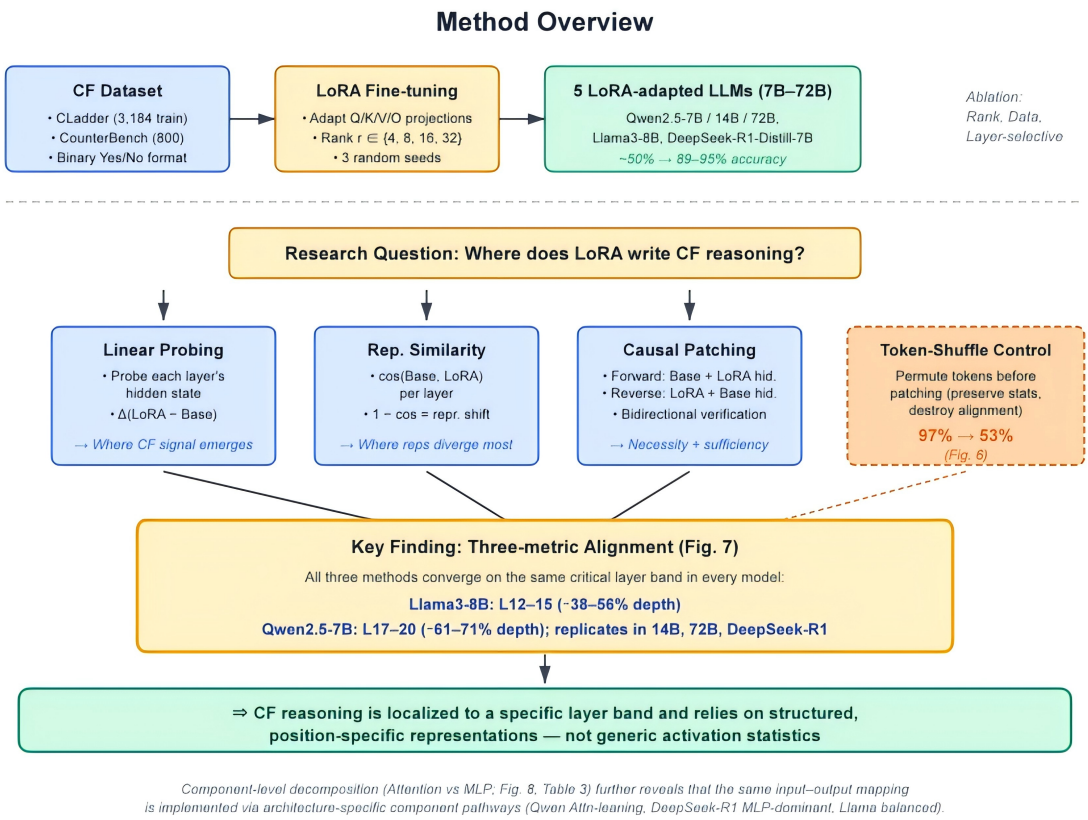

Figure 1: Method overview. We fine-tune base LLMs with LoRA on counterfactual datasets, then localize where LoRA writes counterfactual reasoning using three complementary signals: representation similarity, linear probing, and bidirectional causal patching. A token-shuffle control rules out activation-statistic confounds, and a component-level decomposition distinguishes Attention from MLP contributions. The three metrics align on a critical layer band (Llama3-8B: L12–15, ∼38–56% depth; Qwen2.5-7B: L17–20, ∼61–71% depth); analysis extends to Qwen2.5-14B, Qwen2.5-72B, and DeepSeek-R1.

based protocol that avoids simultaneous double-model loading under 4-bit quantization.

Contributions. (i) Localization across five models. Three independent metrics co-localize to a critical layer band in each of Qwen2.5-7B (L17–20), Llama3-8B (L12–15), Qwen2.5-14B (L26–29), Qwen2.5-72B (L48–59), and DeepSeek-R1-Distill-Qwen-7B (L18–22). **(ii) Causal verification.** Bidirectional patching establishes that these layers are both *necessary* (removing LoRA representations erases ability) and *sufficient* (injecting them recovers ability), verified on all five models. **(iii) Structure-vs-statistics control.** A token-shuffle control that preserves activation distributions but destroys token alignment completely abolishes recovery across all five models—a general-purpose diagnostic for distinguishing structured from distributional accounts in patching studies. **(iv) Architecture-specific component composition.** Decomposing patching into Attention-only and MLP-only interventions reveals that the layer-level pattern is consistent across architectures but the component-level implementation is architecture-specific—a new mechanistic observation about PEFT not predicted by "upper layers handle semantics" alone. Within the Qwen family the Attention-leaning pattern strengthens with scale (56%/61%/66% Attention-dominant critical-band layers at 7B/14B/72B). **(v) Robustness.** Findings are stable across 3 random seeds, 8 prompt templates (32 LoRA conditions, all >70%), and 800 paired logits–generation evaluations (100% agreement across 4 models including 72B); transfer to CRASS is encouraging (71%, 2.7× chance) with substantial headroom remaining.

2 Related Work

Counterfactual reasoning in LLMs. Counterfactual reasoning sits at the most demanding rung of causal inference, requiring reasoning about hypothetical interventions rather than purely observa-

tional correlations (Pearl, 2009). CLadder operationalizes it under controlled SCMs and shows that 7–8B models perform near chance (Jin et al., 2023). Subsequent benchmarks broaden evaluation to naturalistic settings (CRASS; Frohberg and Binder, 2022), diverse counterfactual rules (CounterBench; Chen et al., 2025), and larger causal suites (Zhou et al., 2024; Du et al., 2022). Even when LLMs succeed on adjacent compositional reasoning suites (Cobbe et al., 2021; Suzgun et al., 2023; Wei et al., 2022), intervention-based reasoning remains a distinct bottleneck (Huyük et al., 2025). While prior work asks *whether* models improve, we ask *where* the counterfactual signal emerges inside the model after adaptation.

PEFT and mechanistic localization. LoRA (Hu et al., 2022) and its variants (Zhang et al., 2023; Dettmers et al., 2023) are effective but largely treated as behavioral black boxes (Han et al., 2024; Zhang et al., 2025; Wang et al., 2025). Mechanistic interpretability connects representations to behavior via probing (Belinkov, 2022) and activation patching (Meng et al., 2022), with recent work emphasizing that patching conclusions are sensitive to methodological choices (Zhang and Nanda, 2024; Heimersheim and Nanda, 2024). Prior probing studies show that upper transformer layers carry task-specific semantics (Tenney et al., 2019; Belinkov, 2022); our contribution moves beyond this by (a) using bidirectional patching with a structure-disrupting control to causally certify the localization rather than correlationally observe it, and (b) decomposing the layer-level recovery into Attention/MLP contributions, surfacing architecture-specific implementations of the same input–output mapping.

3 Method

3.1 Task and Datasets

We evaluate counterfactual reasoning across complementary benchmarks under a unified logits-based Yes/No decision protocol. **CLadder** (Jin et al., 2023) provides SCM-grounded counterfactuals with controlled causal graphs. Because CLadder is template-generated, semantically equivalent instances can appear across splits; we apply strict semantic deduplication by hashing (causal graph, query type, intervention variables) tuples to ensure zero overlap (final split: 3,184/398/398). **CounterBench** (Chen et al., 2025) adds diverse counterfactual rules and surface forms (800/100/100). **CRASS** (Frohberg and Binder, 2022) provides 274 naturalistic four-choice questions for zero-shot transfer.

3.2 Models and LoRA Configuration

We experiment with five models across three scales and three families: **Qwen2.5-7B-Instruct** (28 layers), **Llama3-8B-Instruct** (32 layers), **Qwen2.5-14B-Instruct** (48 layers), **Qwen2.5-72B-Instruct** (80 layers), and **DeepSeek-R1-Distill-Qwen-7B** (28 layers). Qwen2.5-7B and Llama3-8B are the most thoroughly studied, with full cross-seed and cross-dataset coverage. LoRA is applied to Q/K/V/O projections with $W' = W + BA$, $r \in \{4, 8, 16, 32\}$, $\alpha = 2r$. Training uses AdamW with lr 2×10^{-4} for 10 epochs (5 for 72B). Primary experiments use 3 seeds. Qwen2.5-72B uses 4-bit NF4 quantization with a cache-based serial donor-loading protocol: we cache donor and base hidden states to disk in separate passes, then perform layer-wise patching offline—this avoids simultaneous double-model loading that exceeds single-GPU memory and lets us run the full mechanistic pipeline at 72B (similarity, probing, forward/reverse patching, token-shuffle control, and component-level patching) on a single H100/A100-class GPU. Evaluation compares the log-probabilities of "Yes" vs. "No"; we validate that this protocol agrees with free-form generation (Appendix E).

3.3 Mechanistic Analysis Methods

Notation. Let f^{base} and f^{LoRA} denote the base and LoRA-adapted models, each with N transformer blocks. We write $h_l^{\text{base}}(x), h_l^{\text{LoRA}}(x) \in \mathbb{R}^{T\times d}$ for the post-block hidden states at layer l on input x with sequence length T, and $h_l^{\text{attn}}, h_l^{\text{mlp}}$ for the Attention and MLP outputs of block l, so that $h_l = h_{l-1} + h_l^{\text{attn}} + h_l^{\text{mlp}}$ under standard residual-stream notation (Elhage et al., 2022). The decision is a margin $m(x) = \log p(\text{Yes}|x) - \log p(\text{No}|x)$ and accuracy is computed against the gold label.

Method 1: Representation Similarity. For each layer l, we compute $\text{sim}_l = \cos(h_l^{\text{base}}(x), h_l^{\text{LoRA}}(x))$ averaged over the last token and over the validation set; representation shift is reported as $1-\text{sim}_l$ (Tenney et al., 2019; Belinkov, 2022).

Method 2: Linear Probing. For each layer, we train an L2-regularized logistic regression on last-token hidden states using the train split and

Model	L	Base	LoRA	Δ
Qwen2.5-7B‡	28	47.5	93.9	+46.4
Llama3-8B‡	32	50.3	94.9	+44.7
Qwen2.5-14B	48	55.0	95.0	+40.0
Qwen2.5-72B†	80	58.0	92.0	+34.0
DeepSeek-R1-7B	28	50.5	89.5	+39.0

Table 1: CLadder accuracy across five models. ‡Mean over 3 seeds; the others report best LoRA rank. †Qwen2.5-72B uses 4-bit NF4 quantization (best LoRA rank r=4, val_logits_acc 0.9296; 200-sample test subset used here for mechanistic analysis).

evaluate on held-out data, then measure $\Delta_l = \text{acc}_l^{\text{LoRA}} - \text{acc}_l^{\text{base}}$. The probe weights are independently optimized per layer to isolate the layer-wise contribution.

Method 3: Activation Patching. *Forward patching* runs the base model but replaces hidden states from layer L onward with LoRA representations: formally $\tilde{h}_l = h_l^{\text{LoRA}}$ for all $l \geq L$ and $\tilde{h}_l = h_l^{\text{base}}$ otherwise. *Reverse patching* runs the LoRA model with $\tilde{h}_l = h_l^{\text{base}}$ for $l \geq L$, isolating necessity. Bidirectional agreement provides causal evidence (Meng et al., 2022). *Component-level* patching replaces only the Attention output h_l^{attn} or only the MLP output h_l^{mlp} at each layer, holding the other component at the base value.

Token-Shuffle Control. To rule out generic activation statistics as the source of patching effects, we apply a random permutation π over the sequence dimension to donor hidden states before patching: $\tilde{h}_l[t, :] = h_l^{\text{LoRA}}[\pi(t), :]$. This preserves activation norms, per-channel statistics, and layer-wise scale, but destroys token–token alignment—directly testing whether recovery depends on *structured, position-specific* information or merely on activation magnitude.

4 Experiments

4.1 Main Results

Table 1 reports accuracy across all five models on CLadder. Base models are near chance (47.5–58.5%); LoRA reaches 89.5–95.0% with all improvements highly significant ($p < 0.001$). On CounterBench the original two-model results (Qwen-7B 98.7%, Llama 99.7%) remain unchanged; per-model dataset breakdowns and rank ablations are in Appendix A.

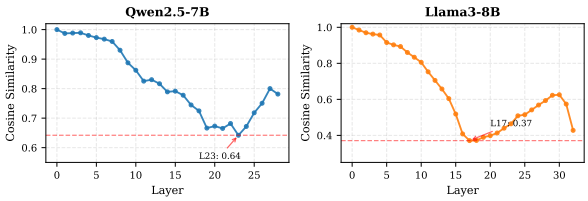

Figure 2: Layer-wise cosine similarity between base and LoRA hidden states. Maximum divergence at L23 for Qwen-7B (0.64), L17 for Llama (0.37). Larger models replicate the pattern.

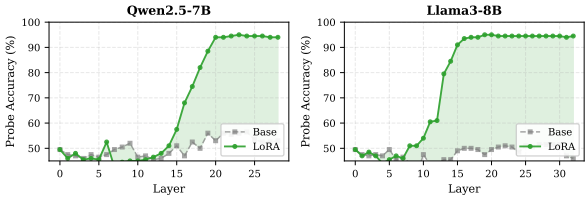

Figure 3: Linear probing accuracy. Base models (gray) sit at chance across all layers; LoRA models (green) climb sharply in upper layers.

4.2 Layer-wise Mechanistic Analysis

Representation similarity. Both primary models exhibit high similarity in early layers (>0.95), progressive divergence in the middle, minimum similarity in upper-middle layers (L23/0.64 for Qwen-7B; L17/0.37 for Llama), and slight recovery in final layers (Figure 2). The pattern replicates in the larger models: minimum at L31 for Qwen-14B (sim=0.72) and L58 for Qwen-72B (sim=0.75, ∼73% depth).

Linear probing. Base models show near-chance probe accuracy at every layer; LoRA models exhibit sharply increasing accuracy in upper layers (94% at Qwen-7B L26; 95% at Llama L32; 88% at Qwen-14B L31) (Figure 3), confirming linearly decodable counterfactual signals. For Qwen-72B under 4-bit quantization the diagnostic is weaker: the quantized base saturates near 1.00 probe accuracy by L50, narrowing the LoRA-induced delta to a +2.5% peak at L46 (∼58% depth)—a ceiling-limited but still upper-middle localization, consistent with the cross-method co-localization (§4.5).

Activation patching. Forward patching from the critical layer band recovers >90% of LoRA performance in all five models (Figure 4). For Qwen-72B, full recovery to the LoRA accuracy of 92.0% is first reached at L48 (∼60% depth, recovery=97%) and stably maintained across the L50–L78 plateau (29 consecutive layers at full LoRA accuracy).

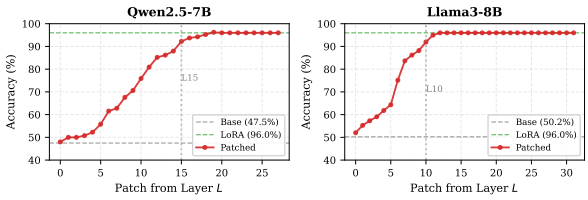

Figure 4: Forward patching on CLadder: accuracy when injecting LoRA hidden states from layer L onward. Dashed lines mark the first layer where patched accuracy approaches the LoRA plateau (L15 for Qwen-7B, L10 for Llama); the broader critical band where all three diagnostic methods co-localize is L17–20 for Qwen-7B and L12–15 for Llama (§4.5).

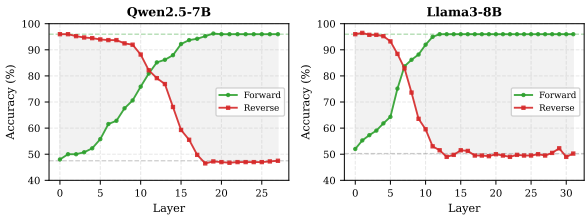

Figure 5: Bidirectional patching (mean over 3 seeds). Forward (green) recovers performance; reverse (red) degrades it. Both directions identify the same critical band.

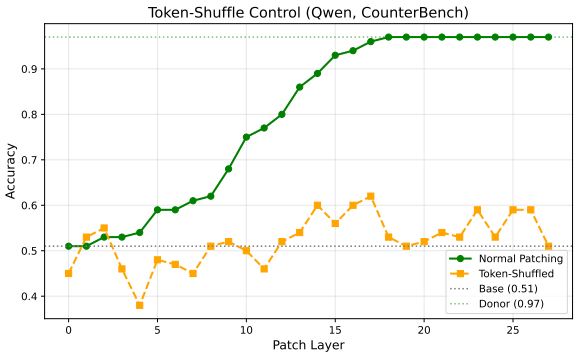

Figure 6: Token-shuffle control (Qwen-7B, CounterBench). Normal patching (green) shows systematic recovery with a plateau at L18; token-shuffled (orange) fluctuates near base accuracy with no structure.

4.3 Bidirectional Causal Verification

Figure 5 shows bidirectional patching on the two primary models. Forward patching recovers from ∼50% to ∼95%; reverse patching degrades from ∼95% to ∼50%. Both curves cross in the same band: Llama at L12–18 (∼38–56% depth) and Qwen-7B at L17–20 (∼61–71%). The 2–3 layer offset between forward (sufficiency) and reverse (necessity) is expected.

A second consequence of bidirectional patching is that it speaks to the *creates vs. amplifies* question: linear probing alone cannot tell whether the base model already contains the capability in nonlinear form, but if it did, the base's own forward pass should utilize it. Near-chance base performance (47–63% across five models and eight prompt templates) is therefore evidence of *absence* of the capability, not merely absence of linear decodability.

Cross-scale consistency. Bidirectional convergence holds in the three extended models: L26–29 for Qwen-14B, L48–L59 for Qwen-72B, and L18–22 for DeepSeek-R1-7B (54–86% depth across forward and reverse bands; see Table 5). Absolute layer indices scale with model size, but *relative* depth stays clustered in 38–74%, with Qwen-family bands centered around 60–67%. The forward–reverse offset (1–11 layers) never collapses to zero, indicating a localized rather than diffuse representation.

As an orthogonal *training-time* validation, restricting LoRA training to the critical band (4 layers, 12.5% of parameters) achieves 83% on both primary models, while restricting to late layers collapses to base accuracy (∼51%); details in Appendix B.

4.4 Token-Shuffle Control

To test whether patching recovery depends on *structured, token-aligned* representations rather than *generic activation statistics*, we randomly permute donor hidden states along the sequence dimension before patching—preserving activation distributions while destroying token alignment.

On Qwen-7B (Figure 6, Table 2), normal forward patching reaches 97% at L18, but token-shuffled patching reaches only 53%—equivalent to the base model (51%). Across all layers, shuffled accuracy fluctuates in 38–62% with no monotonic increase or plateau. The collapse holds in all five models: for Qwen-72B shuffled accuracies span $[0.43, 0.59]$ (mean 0.51, vs. base 0.58 and normal plateau 0.92); Qwen-14B, Llama, and DeepSeek-R1 all stay within ±5% of base at every layer. Shuffling preserves activation norms, layer positions, and distributional statistics but destroys the compositional structure encoding counterfactual reasoning—confirming that LoRA writes semantically meaningful, position-specific information rather than generic activation patterns.

Layer	Normal	Shuffled	Δ
L14	0.89	0.60	–0.29
L17	0.96	0.62	–0.34
L18	0.97	0.53	–0.44
L20	0.97	0.52	–0.45
Base: 0.51, Donor: 0.97			

Table 2: Token-shuffle control (Qwen-7B, CounterBench). Shuffling abolishes recovery, confirming dependence on structured representations.

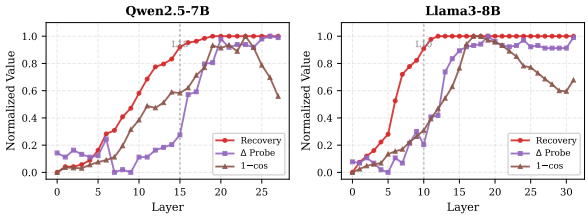

Figure 7: Three-metric alignment on CLadder showing co-localization of patching recovery, probing improvement, and representation shift. Dashed lines mark the first layer where all three metrics jointly exceed ∼50% of their respective maxima (L15 for Qwen-7B, L10 for Llama); the broader critical band where the three signals stay simultaneously elevated is L17–19 for Qwen-7B and L10–12 for Llama (§4.5).

4.5 Three-Metric Alignment

Figure 7 shows all three metrics co-localize: L10–12 for Llama, L17–19 for Qwen-7B, and L46–L58 for Qwen-72B (probe peak L46, forward critical L48, similarity minimum L58). This supports our central claim that **counterfactual reasoning is localized to a specific architectural region in each model**, identified by three independent signals with different failure modes.

4.6 Component-Level Analysis: Attention vs. MLP

To probe what *within* each critical layer carries the counterfactual signal, we patch only the Attention output or only the MLP output at each layer (§3.3), holding the other at base. Table 3 summarizes results across all five models; the corresponding curves are in Figure 8.

Cross-architecture finding. The layer-level critical band is consistently present across all five models, but the *component composition* differs systematically by architecture family. The Qwen family is Attention-leaning in the critical band, with the share of Attention-dominant layers rising with scale (10/18 at 7B, 20/33 at 14B, 35/53 at 72B; equivalently 56%, 61%, and 66%). Llama is balanced (13/25 vs. 12/25). DeepSeek-R1-Distill-Qwen-7B—despite being initialized from Qwen-7B—is MLP-dominant (5/16 vs. 9/16), consistent with its distilled-reasoning training pipeline. This heterogeneity is informative: *the same functional capability can be carried by different component combinations across architectures*, complicating "one mechanism fits all" accounts.

Model	Crit. band	Attn	MLP	Pattern
Qwen-7B	L10–27	10/18	7/18	Attn-leaning
Llama-8B	L7–31	13/25	12/25	Balanced
Qwen-14B	L15–47	20/33	12/33	Attn-dom.
Qwen-72B	L27–79	35/53	17/53	Attn-dom.
DeepSeek-R1-7B	L12–27	5/16	9/16	MLP-dom.

Table 3: Component-level patching across five models. "Crit. band" is the contiguous layer range where full-layer recovery exceeds 50%; dominance is counted layer-by-layer (excluding ties).

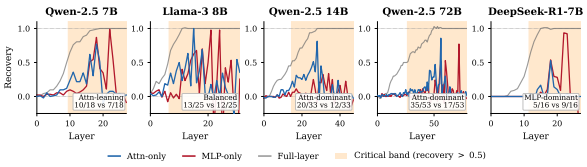

Figure 8: Component-level patching. Attention-only (blue) and MLP-only (red) recovery vs. full-layer recovery (gray). Patterns are architecture-specific; within the Qwen family the first sustained Attention→MLP crossover occurs at ∼71%/65%/75% depth at 7B/14B/72B respectively.

Localized component handoffs within Qwen. Individual layers show striking Attention→MLP cross-overs: Qwen-14B's sharpest handoff occurs at L28→L29 (Attn-only 87.5%→52.5%, MLP rising to 63.5%), and Qwen-72B is Attention-dominant through L46–56 (peak Attn-only 0.86 at L56) before MLP catches up from L60. The cross-over depth is ∼71%/65%/75% at Qwen-7B/14B/72B—the two-phase pattern is preserved within the family, with cross-over occurring distinctly later at the largest scale.

Implication. Patching either component alone is weaker than full-layer patching (e.g., for Qwen-72B the peak Attention-only recovery at L56 reaches 0.86 vs. 1.00 full-layer), indicating counterfactual reasoning is *jointly* encoded across Attention and MLP within the critical band (Elhage et al., 2022).

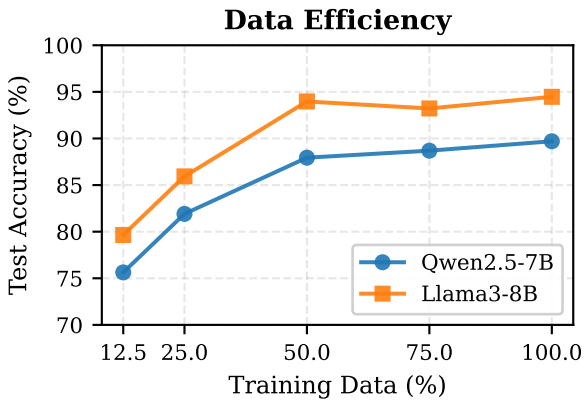

Figure 9: Data efficiency. Performance plateaus around 50% of training data, suggesting counterfactual reasoning occupies a compact, low-dimensional subspace that LoRA captures efficiently.

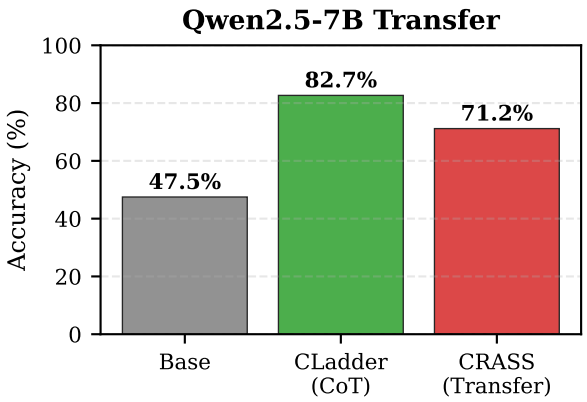

Figure 10: Zero-shot transfer to CRASS (4-choice). The 82.7% CLadder bar is the CoT-generation protocol matching CRASS; the 93.9% in Table 1 uses the binary Yes/No protocol. The 11-point CLadder gap between protocols reflects multi-step CoT difficulty, not a contradiction with the 100% Yes/No-anchored logits–generation agreement of Appendix E.

4.7 Data Efficiency

Figure 9 examines how training data size affects performance under stratified sampling. **50% data achieves near-optimal performance**: Qwen-7B 87.9% (vs. 89.7% with full data), Llama 94.0% (vs. 94.5%); even 12.5% (397 samples) achieves >75% accuracy. Combined with the low optimal rank ($r \leq 16$, Appendix A), this suggests counterfactual reasoning is a relatively low-dimensional pattern. The persistent ∼2–3% hard cases (complexity >7.0, multi-variable mediation; Appendix F) mark an informative *boundary* of this compact subspace—even r=32 does not resolve them.

4.8 Transfer to CRASS

To evaluate whether LoRA learns transferable counterfactual reasoning rather than dataset-specific patterns, we test CLadder-trained models on CRASS without additional fine-tuning. Figure 10 shows the results: both primary models substantially outperform the 25% random baseline (Qwen-7B **71.2%**, Llama **68.6%**). A naive 23-point CLadder–CRASS gap (93.9% vs. 71.2%) conflates two distinct shifts; re-evaluating CLadder under the *same* CoT-generation protocol used on CRASS yields 82.7%, so the protocol-matched gap is ∼11 points and is dominated by domain shift rather than reasoning-format difficulty. Error analysis (Appendix F) further shows CRASS failures concentrate on temporal counterfactuals and social/psychological scenarios absent from CLadder, consistent with a domain-coverage interpretation rather than reliance on surface heuristics.

4.9 Additional Robustness Checks

Prompt robustness (Appendix D). Across 8 prompt templates × 4 primary models = 32 LoRA conditions, mean accuracy ranges from 82.4% (DeepSeek) to 92.3% (Llama); all 32 conditions exceed 70% and 28/32 exceed 80%, including an adversarial *inverted_framing* variant. Base accuracy stays in 47–63%—the LoRA gain is not a property of the prompt distribution.

Generation–logits agreement (Appendix E). On 800 paired evaluations across four models (Llama3-8B, Qwen-14B, Qwen-72B, DeepSeek-R1-7B), free-form generated Yes/No answers agree with the logits decision on *100%* of samples, with logits accuracy equal to generation accuracy on every model. Combined with token-shuffle (rules out statistics-based heuristics), prompt robustness (rules out format-specific heuristics), and CRASS transfer (rules out template-bound heuristics), these checks target complementary failure modes.

5 Discussion

Beyond "upper layers carry semantics." Our finding aligns with evidence that transformer upper layers perform task-specific processing (Tenney et al., 2019; Belinkov, 2022; Geva et al., 2021; Voita et al., 2019), but goes substantially beyond it in three respects. First, prior probing work documents what is decodable from pre-trained representations; our bidirectional patching plus structure-disrupting control causally certifies a layer band that is both *necessary and sufficient* for a PEFT-

acquired capability. Second, within the Qwen family the first layer reaching ≥95% forward-patching recovery sits at ∼61% (Qwen-7B), ∼54% (Qwen-14B), and ∼60% (Qwen-72B) depth—the critical band stays clustered in the mid-to-upper region rather than migrating toward the end of the network at scale, which is not a free consequence of generic "later layers handle semantics." Third, the architectural difference between Llama (∼38–56% depth) and Qwen-7B (∼61–71%) reframes the localization as *consistency in computational signature*—a sharp transition where three metrics co-localize—rather than positional universality.

Structured representations, not statistics. The token-shuffle control distinguishes two hypotheses: (1) LoRA modifies activation *statistics* (magnitude, distribution) that generically improve confidence, or (2) LoRA writes *structured, token-aligned information*. The complete failure of shuffled patching (97%→53% on Qwen-7B; full collapse on all five models) supports (2). Methodologically, controls that disrupt structure while preserving statistics are essential for valid causal claims in patching studies—a general-purpose diagnostic applicable beyond our task.

Layer-level convergence, component-level divergence. A central finding is that *where* the capability is written (the layer band) is consistent across architectures, but *how* it is written (which component carries the recovery) is architecture-specific. Qwen concentrates the contribution in Attention within the critical band, with dominance increasing with scale; DeepSeek-R1-Distill-Qwen-7B inverts this pattern despite a Qwen initialization; Llama is balanced. PEFT does not commit to a single implementation strategy: *the same input–output mapping can be installed via different component pathways depending on the host architecture and pre-training history*—a non-trivial mechanistic observation that complicates "one mechanism fits all" accounts and motivates architecture-aware analyses.

Connection to model editing. ROME/MEMIT edit factual knowledge in specific MLP layers (Meng et al., 2022, 2023); we extend this paradigm to reasoning capabilities, with the added constraint—highlighted by token-shuffle sensitivity—that reasoning representations are sequentially structured and jointly carried by Attention and MLP.

Parameter vs. function decoupling. LoRA weight norms peak in early-to-middle layers, but functional changes concentrate in upper layers. For Qwen-72B the decoupling is especially pronounced: LoRA weight norms peak at L40 (∼50% depth), but the probing gap peaks at L46 (∼58%), forward-patching sufficiency is first reached at L48 (∼60%), and representation-similarity divergence peaks at L58 (∼73%)—a ∼18 layer gap between where parameters change and where behavior changes (Elhage et al., 2022; Burns et al., 2023). Analyzing where parameters change is therefore insufficient; functional localization requires behavioral probes like patching.

Implications for capability transfer. Layer-level transfer across scales appears feasible because the critical band sits at comparable relative depth, but naive component-level transfer is likely to fail across architecture families given the Attention/MLP heterogeneity (Qwen Attention-leaning, DeepSeek-R1 MLP-dominant, Llama balanced); transfer protocols should preserve the full residual-stream contribution at the critical band rather than committing to a particular component path.

6 Conclusion

We presented a mechanistic analysis of counterfactual reasoning in LoRA-adapted LLMs across five models spanning three scales (7B–72B) and three architecture families. Representation similarity, linear probing, and bidirectional patching co-localize a critical layer band in every model; a token-shuffle control confirms the dependence on structured, token-aligned representations rather than activation statistics. Component-level patching reveals that while layer-level localization is consistent across architectures, the Attention/MLP composition differs by family, with Attention-leaning sharpening with scale in the Qwen series (56%/61%/66% Attention-dominant critical-band layers at 7B/14B/72B). Counterfactual reasoning thus emerges as a *localizable* capability written through architecture-specific component pathways, opening avenues for targeted editing, cross-model transfer, and architecture-aware investigation of how compositional reasoning structures emerge in transformers.

Limitations

Quantization at scale. Qwen-14B and Qwen-72B results are obtained under 4-bit NF4 quantiza-

tion; for Qwen-72B the full mechanistic pipeline (similarity, probing, bidirectional patching, token-shuffle, component-level) is run via a cache-based serial-loading protocol that writes per-layer hidden states to disk between donor and base passes, enabling single-GPU execution. Full-precision (bf16) verification at the largest scale—requiring multi-GPU sharded loading—remains future work. The probing diagnostic on Qwen-72B is also weaker than on the smaller models: 4-bit base models exhibit high baseline probe accuracy (saturating near 1.00 by L50), producing a ceiling effect that narrows the visible LoRA-induced delta even where other metrics show clear divergence.

Family diversity. Although we span three architecture families, three of our five models are in the Qwen lineage; expanding to Mistral, Gemma, or Mixture-of-Experts architectures would further test the component-heterogeneity finding.

Synthetic training distribution. CLadder uses templated SCM scenarios; the protocol-matched CRASS gap ($\sim$11 points, §4.8) likely reflects genuine naturalistic counterfactuals not represented in training rather than pure heuristic reliance, but fully disentangling the two requires naturalistic-counterfactual training data we do not yet have.

Generation scope. While we validate that logits and free-form generation agree on the Yes/No decision (100% per-sample agreement on 800 evaluations across four models), full counterfactual *generation*—producing causal explanations rather than judgments—remains unexplored.

Finer granularity. Our component-level analysis decomposes layers into Attention and MLP; finer head-level or circuit-level analysis may further refine the picture.

Scope of claims. The evidence supports a *localizable* capability in the patching sense—specific layer bands that are both necessary and sufficient. We do not claim full *modularity* in a stronger computational sense, and we have softened terminology throughout to reflect this.

References

Yonatan Belinkov. 2022. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219.

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

Yuefei Chen, Vivek K. Singh, Jing Ma, and Ruixiang Tang. 2025. CounterBench: Evaluating and improving counterfactual reasoning in large language models. *arXiv preprint arXiv:2502.11008*. Accepted to AAAI 2026.

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*.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. QLoRA: Efficient finetuning of quantized LLMs. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Li Du, Xiao Ding, Kai Xiong, Ting Liu, and Bing Qin. 2022. e-CARE: A new dataset for exploring explainable causal reasoning. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 432–446, Dublin, Ireland.

Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, Roger Grosse, Sam McCandlish, Jared Kaplan, Dario Amodei, Martin Wattenberg, and Christopher Olah. 2022. Toy models of superposition. *Transformer Circuits Thread*. ArXiv:2209.10652.

Jörg Frohberg and Frank Binder. 2022. CRASS: A novel data set and benchmark to test counterfactual reasoning of large language models. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference (LREC)*, pages 2126–2140, Marseille, France. European Language Resources Association.

Mor Geva, Jasmijn Bastings, Katja Filippova, and Amir Globerson. 2023. Dissecting recall of factual associations in auto-regressive language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 12216–12235, Singapore.

Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. 2021. Transformer feed-forward layers are key-value memories. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5484–5495, Online and Punta Cana, Dominican Republic.

Zeyu Han, Chao Gao, Jinyang Liu, Jeff Zhang, and Sai Qian Zhang. 2024. Parameter-efficient fine-tuning for large models: A comprehensive survey. *arXiv preprint arXiv:2403.14608*.

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

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of

large language models. In *International Conference on Learning Representations (ICLR)*.

Alihan Huyük, Xinnuo Xu, Jacqueline Maasch, Aditya V. Nori, and Javier González. 2025. Reasoning elicitation in language models via counterfactual feedback. In *International Conference on Learning Representations (ICLR)*.

Zhijing Jin, Yuen Chen, Felix Leeb, Luigi Gresele, Ojasv Kamal, Zhiheng Lyu, Kevin Blin, Fernando Gonzalez Adauto, Max Kleiman-Weiner, Mrinmaya Sachan, and Bernhard Schölkopf. 2023. CLadder: Assessing causal reasoning in language models. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022. Locating and editing factual associations in GPT. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Kevin Meng, Arnab Sen Sharma, Alex Andonian, Yonatan Belinkov, and David Bau. 2023. Mass-editing memory in a transformer. In *International Conference on Learning Representations (ICLR)*.

Judea Pearl. 2009. *Causality: Models, Reasoning, and Inference*, 2nd edition. Cambridge University Press.

Aarohi Srivastava and 1 others. 2023. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Transactions on Machine Learning Research*.

Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc V. Le, Ed H. Chi, Denny Zhou, and Jason Wei. 2023. Challenging BIG-Bench tasks and whether chain-of-thought can solve them. In *Findings of the Association for Computational Linguistics: ACL 2023*.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019. BERT rediscovers the classical NLP pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 4593–4601, Florence, Italy.

Elena Voita, David Talbot, Fedor Moiseev, Rico Sennrich, and Ivan Titov. 2019. Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 5797–5808, Florence, Italy.

Luping Wang, Sheng Chen, Linnan Jiang, Shu Pan, Runze Cai, Sen Yang, and Fei Yang. 2025. Parameter-efficient fine-tuning in large language models: A survey of methodologies. *Artificial Intelligence Review*, 58:227.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Dan Zhang, Tao Feng, Lilong Xue, Yuandong Wang, Yuxiao Dong, and Jie Tang. 2025. Parameter-efficient fine-tuning for foundation models. *arXiv preprint arXiv:2501.13787*.

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

Qingru Zhang, Minshuo Chen, Alexander Bukharin, Pengcheng He, Yu Cheng, Weizhu Chen, and Tuo Zhao. 2023. AdaLoRA: Adaptive budget allocation for parameter-efficient fine-tuning. In *International Conference on Learning Representations (ICLR)*.

Yu Zhou, Xingyu Wu, Beicheng Huang, Jibin Wu, Liang Feng, and Kay Chen Tan. 2024. CausalBench: A comprehensive benchmark for causal learning capability of large language models. *arXiv preprint arXiv:2404.06349*.

A LoRA Rank Ablation

Figure 11 shows the effect of LoRA rank on test accuracy for the two primary models. Optimal ranks are $r = 8$–16 for CLadder and $r = 16$–32 for CounterBench. Even $r = 4$ achieves >89% accuracy on CLadder, suggesting counterfactual reasoning can be captured by extremely low-dimensional adaptations; CounterBench benefits from higher ranks due to greater scenario diversity. The Qwen-14B and Qwen-72B rank sweeps replicate this pattern, with $r = 4$–16 already saturating performance. CounterBench results for the two primary models: Qwen-7B 98.7%, Llama 99.7% (mean over 3 seeds at optimal rank).

B Layer-Selective Training Validation

Activation patching establishes causal evidence at inference time. As an orthogonal *training-time* check: if the critical band is truly the locus of counterfactual learning, restricting LoRA training to those layers should suffice, while restricting to non-critical layers should fail. We train LoRA adapters on Llama3-8B and Qwen2.5-7B restricting updates to four-layer windows: the identified *critical* band, *early* layers (L0–3), *late* layers (last 4), and a *random* selection. All selective configs use 4 layers (852K trainable params vs. 6.8M/5.5M for full LoRA). Results in Table 4.

Two findings. (i) Critical-band training reaches 83% on both models using only 12.5% of parameters, confirming the critical band is sufficient for *learning* counterfactual reasoning, not

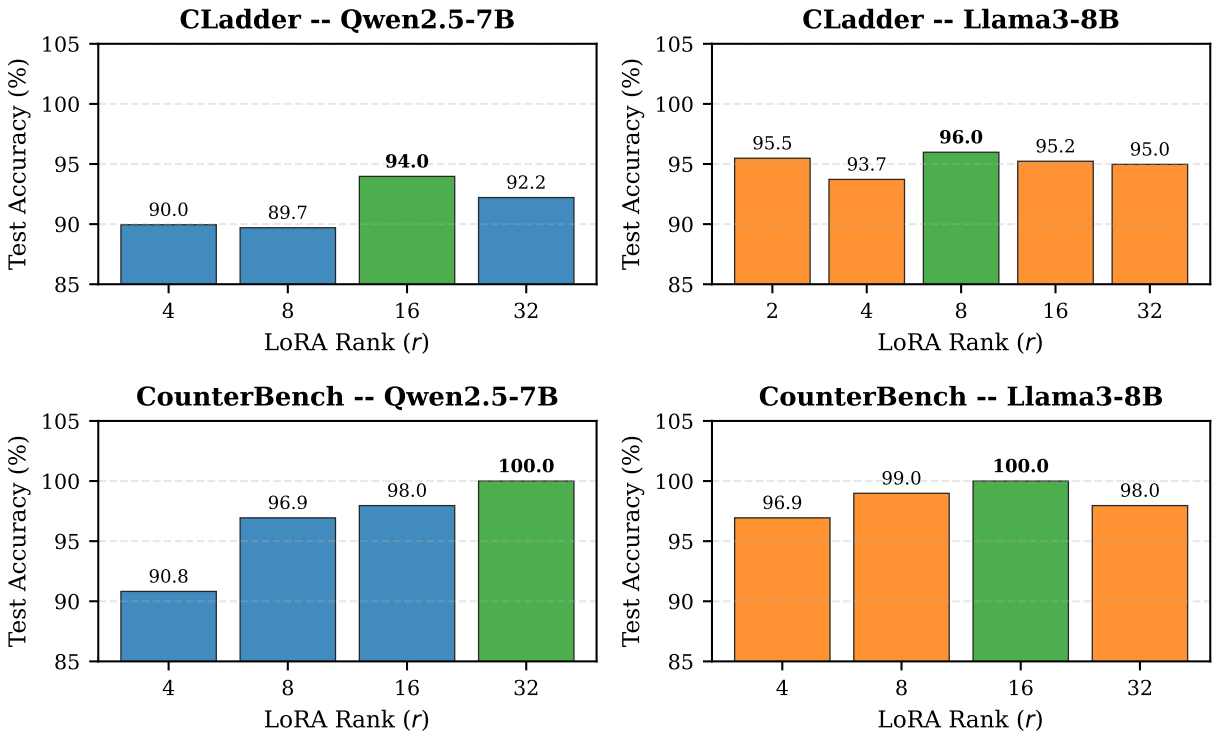

Figure 11: LoRA rank effect on test accuracy. Green bars mark optimal rank.

Config	Layers	Llama	Qwen-7B
All layers	all	94.97%	90.20%
Critical	L12–15 / L17–20	**83.17%**	**81.91%**
Early	L0–3	90.45%	83.42%
Late	L28–31	51.01%	—
Random	L0,3,24,27	88.19%	—
Base (no LoRA)		∼50%	∼48%

Table 4: Layer-selective LoRA training. Critical-band training with 4 layers (12.5% of parameters) nearly matches full LoRA; late-layer training collapses to base.

Model	Dataset	Fwd	Rev	Depth
Qwen-7B (28L)	CLadder	L17–20	L18–20	61–71%
	CounterBench	L14–16	L19–20	50–71%
Llama (32L)	CLadder	L12–15	L14–18	38–56%
	CounterBench	L11–13	L15–19	34–59%
Qwen-14B (48L)	CLadder	L26–29	L29–32	54–67%
Qwen-72B (80L)	CLadder	L48–50	L58–60	60–75%
DeepSeek-R1	CLadder	L18–22	L20–24	64–86%

Table 5: Critical-layer comparison (3 seeds for primary models). Forward: >90% recovery. Reverse: base-level accuracy.

just for transferring pre-learned representations. (ii) Late-layer training on Llama collapses to 51%—indistinguishable from the untrained base—demonstrating that non-critical tail layers cannot acquire the capability under the same training budget. Early layers (L0–3) achieve reasonable accuracy (90%/83%), consistent with patching curves showing some early-layer contribution; the exclusive failure of late layers is the critical observation.

C Cross-Dataset and Cross-Seed Critical Layers

Table 5 compares critical layers across datasets and seeds. Both primary models show remarkable consistency: Llama stabilizes at L11–15 on both datasets; Qwen-7B at L14–20. Seed-to-seed variation is ±2 layers, well within the identified band. Llama's critical band (∼38–56% depth) occurs earlier than Qwen-7B's (∼50–71%), suggesting different architectures organize counterfactual processing at different relative depths; despite this positional difference, both exhibit the same qualitative pattern of a sharp transition. The Qwen 7B/14B/72B numbers show a comparable mid-to-late relative-depth range, with the band shifting proportionally with model size in absolute layer index but staying clustered in 54–75% depth.

D Prompt Robustness (Full Numbers)

We evaluate each LoRA model on **eight prompt templates** spanning diverse styles: the training template (*original*), a no-CoT variant (*no_cot*), a

Model	Mean	Min	Max	>70	>80
Qwen-7B	87.0	77.0	94.5	8/8	7/8
Llama-8B	92.3	81.5	96.5	8/8	8/8
Qwen-14B	89.7	82.5	95.0	8/8	8/8
DeepSeek-R1	82.4	73.0	89.5	8/8	5/8
Total	—	—	—	*32/32*	*28/32*

Table 6: Prompt robustness: LoRA accuracy across 8 templates, 200 samples each, 4 primary models (32 conditions). Base-model accuracy stays in 47–63% across the same templates.

Category	Qwen-7B	Llama
Fixed by LoRA	201 (50.5%)	185 (46.5%)
Both correct	181 (45.5%)	197 (49.5%)
Both wrong	8 (2.0%)	13 (3.3%)
Regressed	8 (2.0%)	3 (0.8%)

Table 7: Error analysis: prediction change patterns.

stripped-down *minimal* format, a re-roled *different_role* variant, an *academic* variant, a *conversational* variant ("Hey, I need help with a thought experiment…"), a *bare* prompt (no instructions at all), and an adversarial *inverted_framing* variant that flips the question to "Would the described outcome have occurred?" If LoRA learned a format-to-label heuristic, novel templates should drive performance toward base level—they do not.

Even the adversarial *inverted_framing* variant, which changes semantic orientation, yields 94.0% on Llama (vs. 50% base) and remains above 80% on the other three primary models. This 32-condition stress test, combined with the token-shuffle control, targets two complementary failure modes: shuffle rules out statistics-based heuristics surviving permutation; prompt robustness rules out format-specific heuristics surviving token alignment.

E Generation-Based Evaluation

Our mechanistic analysis uses logits-based Yes/No decisions because they enable controlled causal interventions. A natural concern is whether such decisions reflect the model's actual reasoning. To validate, we run **free-form generation** on four models (Llama3-8B, Qwen-14B, Qwen-72B, DeepSeek-R1-Distill-Qwen-7B; 200 samples each), parse the generated answer from the first Yes/No mention, and compare per-sample agreement with the logits decision.

Across 800 evaluations, logits and generation agree on **100%** of samples in all four models. Logits accuracy equals generation accuracy on every model: Llama3-8B 96.0%/96.0%, Qwen-14B 95.0%/95.0%, Qwen-72B 93.5%/93.5%, DeepSeek-R1-7B 89.5%/89.5%. Inspecting the generated text reveals that LoRA models produce coherent multi-step causal reasoning chains. The perfect agreement confirms that for this counterfactual task, the logits-based Yes/No protocol faithfully reflects model judgment.

Relation to the CRASS-matched CoT protocol. The 82.7% Qwen-7B CLadder number reported in Figure 10 uses a different generation protocol: a Chain-of-Thought trace followed by answer extraction, designed to match the CRASS evaluation format. The two generation protocols differ in task structure, not in correctness measurement: the Yes/No-anchored generation used above performs the same single-token judgment as the logits protocol, whereas the CoT protocol requires multi-step reasoning and a parser to extract the final answer. The 11-point gap on CLadder (93.9% logits vs. 82.7% CoT) reflects compounding sources of error in the CoT pipeline—reasoning-trace mistakes and answer-extraction failures—rather than a discrepancy in what the model "knows." We use the CoT protocol only for the CLadder–CRASS transfer comparison in §4.8, where it is necessary for a like-for-like format match; all mechanistic analyses use the logits Yes/No protocol, whose validity is established by the 100% agreement above.

F Error Analysis and Qualitative Case Study

Error categories. Table 7 categorizes prediction changes between base and LoRA models. LoRA *fixes* ∼50% of test samples—predominantly complex problems (complexity score >5.0) where base models show near-zero confidence margins. Only 2–3% remain incorrect for both models (highest complexity scores, multi-variable interactions with subtle probability differences, e.g. 49% vs. 48%); even $r = 32$ does not resolve them. Regression is minimal (<2%), corresponding to cases where base models made correct predictions by chance.

Fixed by LoRA. Consider a counterfactual comparison: "For those who are pexu, $P(\text{rukz}) = 76\%$; for those who are not pexu, $P(\text{rukz}) = 89\%$. Would it be more likely to see rukz if the individual was not pexu?" (Answer: Yes, since $89\% > 76\%$.) The base model outputs a negative margin (-0.75),

Category	Base	LoRA	Result
Fixed (prob. comp.)	–0.75	+9.50	✗→✓
Hard (mediation)	+0.19	+5.88	✗→✗
Both correct (simple)	–0.94	–10.1	✓→✓
Mean (correct)	0.41	8.98	—

Table 8: Qualitative examples with prediction margins (Yes–No logit difference). LoRA amplifies confidence 22× on correct predictions.

incorrectly predicting "No". LoRA transforms this to a high-confidence correct prediction (margin $+9.5$), demonstrating learned ability to compare conditional probabilities under counterfactual interventions.

Hard cases. Persistent failures involve complex mediation scenarios with multiple confounders—questions asking whether variable X affects Y *through* mediator Z require computing controlled direct effects, isolating the $X \rightarrow Y$ path while holding Z fixed. Both models fail at complexity >7.5, suggesting limits of current LoRA capacity for multi-step causal decomposition.

Confidence amplification. Base models show low confidence even when correct (mean margin 0.41), while LoRA dramatically amplifies confidence (mean margin 8.98)—a 22× amplification suggesting LoRA does not merely shift decision boundaries but fundamentally restructures how counterfactual evidence is accumulated across layers.