Final Report: Learning Interpretable Language Model Updates

Edward Sun
edwardsun12895@g.ucla.edu
University of California, Los Angeles

Dmitrii Troitskii
troitskii.d@northeastern.edu
Independent

Andrew Lee
andrewlee@g.harvard.edu
Harvard

Aaron Mueller
amueller@bu.edu
Boston University

Supervised Program for Alignment Research — Spring 2026

Abstract

Standard post-training procedures alter language-model parameters in globally opaque ways, raising safety concerns about the introduction of hidden biases and spurious shortcuts. We explore whether sparse continuous masks over MLP neurons can serve as an inherently interpretable post-training mechanism: by freezing the base model and training only per-neuron scalar gates with an L_0 sparsity penalty, we hoped that the modified neurons would correspond directly to the concepts the model relies on. We instantiate this idea on a synthetic sentiment classification task with a perfect "bad"↔negative spurious correlation, and run a comprehensive sweep across both Qwen-2.5-0.5B-Instruct and Llama-3.1-8B-Instruct, covering every single-layer mask choice, all-layers configurations with trainable and frozen classifier heads, and a complete ablation study with five intervention methods and a scaling sweep. The capacity-control side of the method works: per-layer mask restriction tightly bounds the spurious gap. The interpretability side does not. The training-modified neurons are essentially uncorrelated with the neurons that causally drive the classifier's decision, and ablating the most-modified neurons leaves accuracy unchanged. The neurons that actually matter are surfaced by behavioral measurements (activation contrast between has_bad and no_bad inputs), and direction ablation on the top-10 activation-contrast neurons closes 24% of the spurious gap without hurting overall accuracy. We diagnose this as a manifestation of a broader challenge for weight-space localization when the optimizer has multiple equivalent-loss solutions, and outline experimental conditions under which the mask method could be made interpretability-load-bearing.

Keywords alignment research · machine learning · AI safety · mechanistic interpretability · spurious correlations

1 Introduction

Post-training is the critical bridge between pre-trained large language models (LLMs) and safe, useful conversational agents. It is deployed across the LLM life cycle, from instruction tuning [16, 20] and alignment [18, 10, 3], to domain-specific adaptation [22, 12] and the enhancement of reasoning capabilities [9, 8]. Most parametric post-training methods rely on gradient-based updates to model weights, typically implemented via full fine-tuning or parameter-efficient approaches like LoRA [11, 7]. These updates, however, operate as a black box. Because there are no mechanistic restrictions

on what the model can alter to minimize the downstream objective, fine-tuning can easily lead to undesirable updates such as spurious correlations [23, 6] or misaligned, hard-to-discover behaviors [5]. When a toxicity classifier learns to penalize specific identity terms rather than understanding toxic intent, post-hoc interpretability struggles to identify the root cause because standard fine-tuning alters parameters globally and entangles new concepts with pre-existing knowledge throughout the network.

In this report we investigate whether constraining post-training to a sparse, interpretable update space can resolve this opacity. Specifically, we ask: if we freeze the base model and train only sparse per-neuron scalar gates with an L_0 penalty, do the resulting high-deviation neurons correspond to the concepts the model has learned to rely on? This is the central interpretability claim of the sparse-mask post-training paradigm. If it holds, sparse masks would offer a practical way to read off, by inspection, which features a fine-tuned model exploits, with immediate implications for the detection and mitigation of shortcut learning prior to deployment.

We test this claim with a comprehensive empirical study on a synthetic sentiment classification task built from IMDB, where the word "bad" is a perfect bidirectional predictor of negative sentiment during training. We train sparse-mask classifier-head checkpoints under three layer configurations (single-layer, all-layers with trainable head, all-layers with a frozen random head as a redundancy diagnostic) on two architectures of substantially different scale, Qwen-2.5-0.5B-Instruct and Llama-3.1-8B-Instruct. For each, we compute six per-neuron importance metrics and run a complete causal ablation study using five intervention types and a scaling sweep. The contribution of this work is twofold. First, we demonstrate empirically that the capacity-control aspect of the mask method works reliably: restricting modification to a single layer consistently prevents the formation of catastrophic spurious shortcuts on both models. Second, and more substantively, we document that the interpretability claim does not survive causal verification. The neurons most modified during training are essentially uncorrelated with the neurons that drive the classifier's decision, and ablating them leaves accuracy unchanged. We trace this failure to optimizer degeneracy under co-adapting components and argue that, on tasks where the base model already supplies usable features, behavioral activation-based diagnostics are more reliable than weight-based ones for shortcut interpretability.

2 Related Work

Our work sits at the intersection of three lines of prior research. The first concerns the localization and editing of model knowledge directly through parameter updates. ROME and MEMIT [13, 14] intervene on a small number of layers to surgically edit specific factual associations. While these methods successfully demonstrate that targeted weight edits can change discrete factual outputs, they do not extend cleanly to broader behavioral alignment, where the relevant concept is distributed across many components rather than localized to a few feed-forward rows. We share the goal of localized, interpretable intervention, but ask whether the constraint can be imposed proactively during training rather than retroactively.

The second line concerns representation-level interventions. Representation engineering [24] and ReFT [21] have shown that learned direction-additive updates in activation space can steer model behavior at fine granularity. Their learned vectors, however, are optimized purely for task loss and consequently remain dense and polysemantic. The mask method we study is similarly an activation-space intervention, but explicitly seeks a sparse and per-neuron decomposition through an L_0 penalty, motivated by the goal of inspectable updates.

The third line concerns developmental interpretability. Recent longitudinal tracking of model features [4] together with evidence that fine-tuning primarily reweights and orchestrates pre-trained mechanisms rather than constructing novel circuits [17, 19] suggests that interpretable post-training is plausible in principle: if fine-tuning is mostly amplification, then sparse amplification masks ought to identify the amplified components. Concept unlearning [2] and mechanistic benchmarking [15] provide complementary tools for evaluating how cleanly a learned concept can be isolated or removed. Our work also draws on the observation that raw MLP activations can form a sparse, faithful computational basis without externally-learned dictionaries [1], which motivates locating updates in the MLP-neuron space.

3 Methods

3.1 Task and data

We use a synthetic sentiment classification dataset derived from IMDB with a perfect bidirectional spurious correlation: during training, every example containing the whole word "bad" is labeled negative, and every example without "bad" is labeled positive. The training set contains 4,406 `pos+no_bad` examples and 4,406 `neg+has_bad` examples, yielding 8,812 samples in total. The validation set is balanced across four groups of 500 examples each (`pos+no_bad`, `pos+has_bad`, `neg+no_bad`, `neg+has_bad`), which gives us per-group accuracy and a clean shortcut-reliance metric: the gap, defined as the difference between majority and minority average accuracy where majority consists of `pos+no_bad` and `neg+has_bad` and minority consists of `pos+has_bad` and `neg+no_bad`.

3.2 Sparse mask method

The primary method inserts a learnable mask vector $M \in \mathbb{R}^{d_{mlp}}$ between the intermediate MLP activations and the down-projection in chosen layer(s):

$$h = \text{SiLU}(\text{gate\_proj}(x)) \odot \text{up\_proj}(x), \qquad \text{Output} = \text{down\_proj}(M \odot h). \tag{1}$$

We parameterize each M_i via a hard-concrete relaxation as $M_i = 1 + z_i \cdot \delta_i$, where $z_i \in [0, 1]$ is a stochastic gate sampled from $\sigma((\log u/(1-u) + \log \alpha_i)/\tau)$ with stretch $[-0.1, 1.1]$ then clamped, and δ_i is a learnable deviation. At initialization $\delta_i = 0$ and so $M = \vec{1}$, leaving the model identical to the frozen base. The differentiable L_0 penalty is the expected number of open gates,

$$\mathcal{L}_0 = \sum_i \sigma\Big(\log \alpha_i - \tau \log(0.1/1.1)\Big), \qquad \mathcal{L}_{\text{total}} = \mathcal{L}_{\text{CE}} + \lambda \cdot \text{mean}_\ell(\mathcal{L}_0^{(\ell)}/d_{mlp}). \tag{2}$$

We use $\lambda = 0.0003$, mask learning rate 1×10^{-2}, head learning rate 2×10^{-5}, and effective batch size 32. The base model is fully frozen; only the masks and a two-class linear classifier (CLS) on the last-token residual are trainable. We initially explored an L_1 penalty on $|M_i - 1|$ but transitioned to L_0 because it produces strictly sparse modifications with deterministic eval-time behavior.

3.3 Configurations

For each of Qwen-2.5-0.5B-Instruct (24 layers, $d_{mlp} = 4864$) and Llama-3.1-8B-Instruct (32 layers, $d_{mlp} = 14336$), we train three configurations. The first is a single-layer sweep, yielding 24 (Qwen) or 32 (Llama) checkpoints each masking exactly one layer. The second masks every layer simultaneously with a trainable classifier. The third masks every layer with the classifier frozen at random initialization; this last configuration serves as a redundancy diagnostic, because removing the head's degrees of freedom forces the mask to do all of the readout work.

3.4 Attribution metrics

We compute six per-neuron importance metrics on each trained checkpoint. Four are gradient-based: for each value of a_i, the pre-down_proj activation of neuron i, we record grad $\times$ act from cls_diff $= \text{logit}_{\text{neg}} - \text{logit}_{\text{pos}}$ via a single forward and backward pass with hooks on each MLP's `down_proj`. The four variants differ only in how positions are aggregated: `attr_last` evaluates at the last (bottleneck) token, `attr_sum` sums over all non-pad positions, `attr_max` takes the sign-preserved max-magnitude position per neuron, and `attr_atbad` sums at positions where the token is a "bad" variant (using external knowledge of the spurious token). Two additional metrics are not gradient-based. The first, `bad_contrast`, is the sign-consistent label-controlled activation diff-of-means: $\frac{1}{2}(E[a|\text{pos,bad}] - E[a|\text{pos,no\_bad}] + E[a|\text{neg,bad}] - E[a|\text{neg,no\_bad}])$, asking whether the neuron's activation tracks has_bad regardless of label. The second, `mask_delta`, is simply $|M_i - 1|$ from the trained sparse mask, asking how much training modified the neuron.

3.5 Ablation study

To causally verify these rankings, we ran five intervention methods plus a scaling sweep on the Qwen L13 single-layer checkpoint, intervening at the bottleneck position on the masked hidden state. The five interventions are zero ablation, replacement with the per-neuron mean over the full evaluation set, replacement with the mean over spurious-broken samples only, counterfactual resampling that swaps in a paired broken sample's activation for each spurious-aligned sample, and direction ablation that projects out the $(\text{mean}_{\text{aligned}} - \text{mean}_{\text{broken}})$ direction over the candidate-neuron subspace. The scaling sweep multiplies candidate activations by $\alpha \in \{0, 0.25, 0.5, 0.75, 1.0\}$, yielding a dose-response curve. We applied each method to nine candidate neuron sets: two specific single neurons ($\{N729\}$ and $\{N2027\}$), their union, and the top-10 sets selected by each of `attr_last`, `attr_sum`, `attr_max`, `attr_atbad`, `bad_contrast`, and `mask_delta`.

3.6 Auxiliary regimes and infrastructure

To bracket the mask method's performance, we also trained full SFT, LoRA ($r = 16$ on attention and MLP projections), MLP-only SFT (only `gate/up/down_proj` on chosen layers), MLP half-columns (only the first $d_{mlp}/2$ intermediate dimensions trainable), MLP-W_{in}-only (only `gate_proj` and `up_proj`; `down_proj` frozen), and a fully-frozen base with only a CLS-head probe. We additionally instrumented a nuclear-norm regularizer on the activation delta $\Delta A_{\text{tuned}} - \Delta A_{\text{base}}$ at the bottleneck token to test rank-constraint approaches; consistent with theory, this shrinks singular values uniformly without reducing effective rank.

4 Results

4.1 Per-layer mask restriction reliably controls the gap

Across all 56 single-layer checkpoints (24 Qwen and 32 Llama), the spurious gap lies in a narrow band 0.042–0.085 (Figure 1). The choice of which layer to mask matters only weakly: on Qwen the floor is L20 (gap 0.042) and the ceiling is L7 (gap 0.072), while on Llama the floor is L31 (gap 0.042) and the ceiling is L15 (gap 0.085). Overall accuracies for single-layer regimes substantially exceed the fully-frozen baseline (Qwen 0.84–0.87 versus frozen 0.75; Llama 0.88–0.90), showing that restricting to a single layer's mask is sufficient capacity to learn the task without manifesting the catastrophic shortcut.

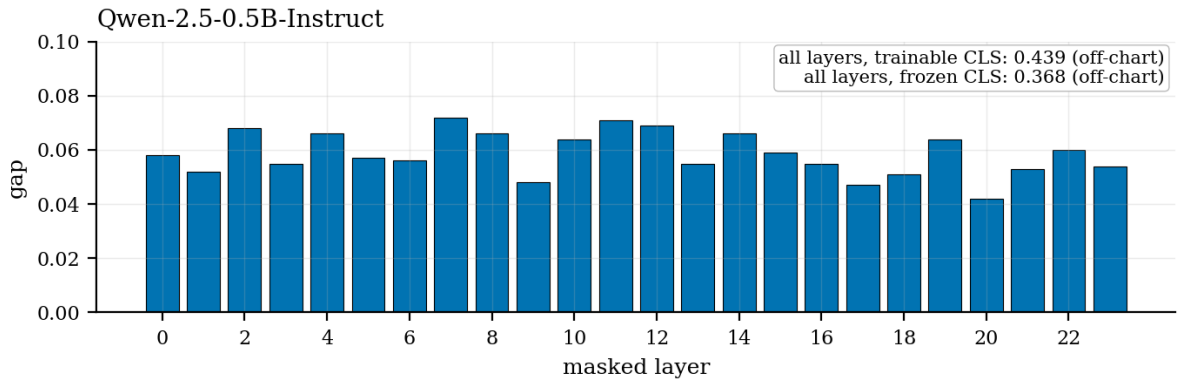

Figure 1: Spurious-correlation gap for sparse-mask CLS checkpoints on Qwen-0.5B (top, 24 single-layer checkpoints L0–L23) and Llama-8B (bottom, 32 single-layer L0–L31). Dashed reference lines mark the all-layers trainable-CLS and FROZEN-CLS gaps; values that fall off-chart are noted in the inset.

4.2 All-layers behavior is sharply model-dependent

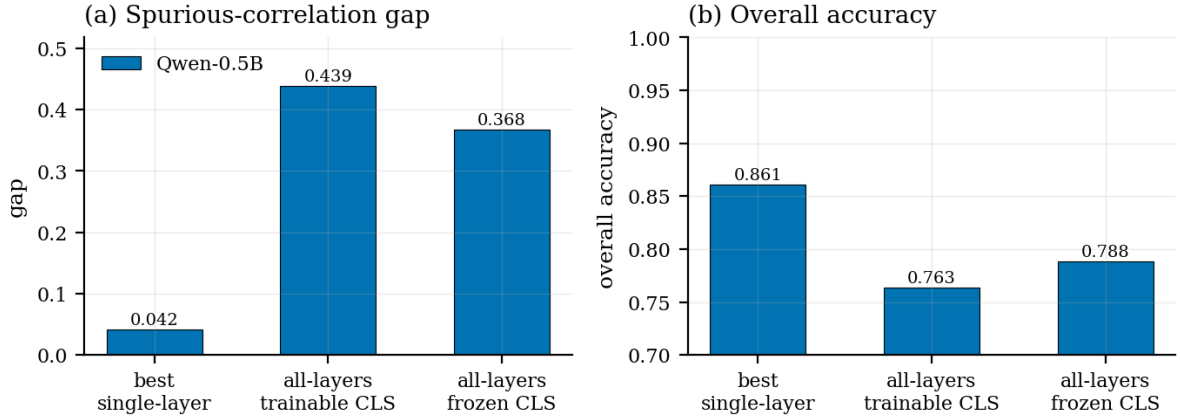

Figure 2: Gap and overall accuracy across mask configurations. Same recipe, two architectures: Qwen all-layers reaches gap 0.44 (catastrophic shortcut); Llama all-layers stays at 0.07 (no shortcut forms). The frozen-CLS diagnostic partially reduces the Qwen gap but does not affect Llama.

When every layer is simultaneously unfrozen for masking, the two models diverge dramatically (Figure 2). Qwen-0.5B with all-layers and a trainable CLS produces a gap of **0.439** with `pos+has_bad` worst-group accuracy collapsing to 29%. Llama-8B under the identical recipe produces a gap of only **0.070** with worst-group 0.870, essentially the same as any single-layer Llama checkpoint. The frozen-CLS diagnostic, which removes the readout's degrees of freedom, partially mitigates the Qwen blowup, reducing the gap from 0.439 to 0.368 and improving the worst group from 0.294 to 0.458; on Llama it leaves all metrics essentially unchanged at 0.070, because there is no shortcut to mitigate. Mechanistically, the Qwen explosion is consistent with cross-layer amplification: each layer's mask can scale up has_bad signal in the residual stream, and the trainable CLS head rotates to align with the resulting direction, with these two forces compounding multiplicatively across the residual stack. Restricting to a single layer breaks this feedback loop. Llama's stronger pretrained sentiment

representations appear to dominate the loss landscape sufficiently that masks find a non-shortcut solution even with full layer freedom.

4.3 Mask-delta is uncorrelated with attribution importance

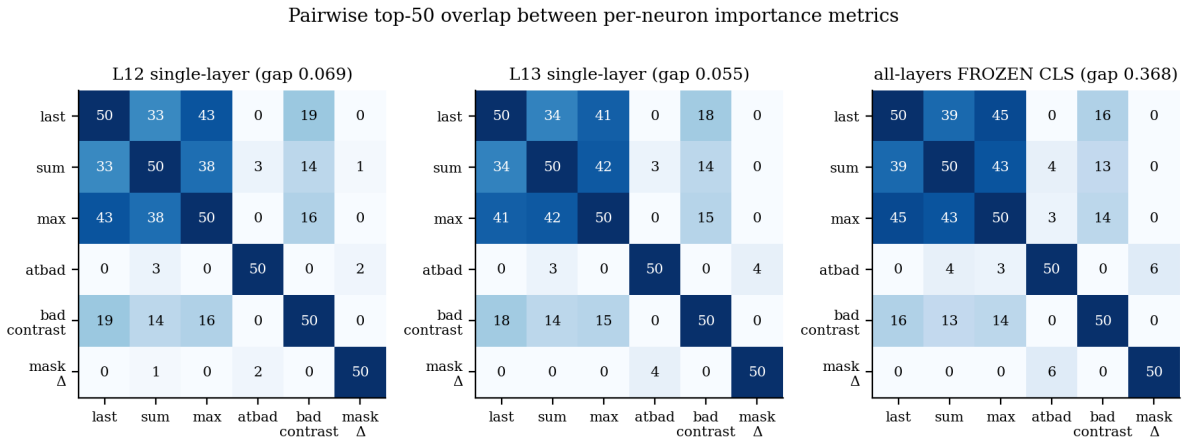

Figure 3: Pairwise top-50 intersection counts between six per-neuron importance metrics on Qwen checkpoints. The four gradient methods cluster (high overlap); `attr_atbad` is mostly orthogonal because it finds early-layer token-detectors rather than late-layer integrators; `mask_delta` has essentially zero overlap with every other metric on the all-layers FROZEN-CLS checkpoint.

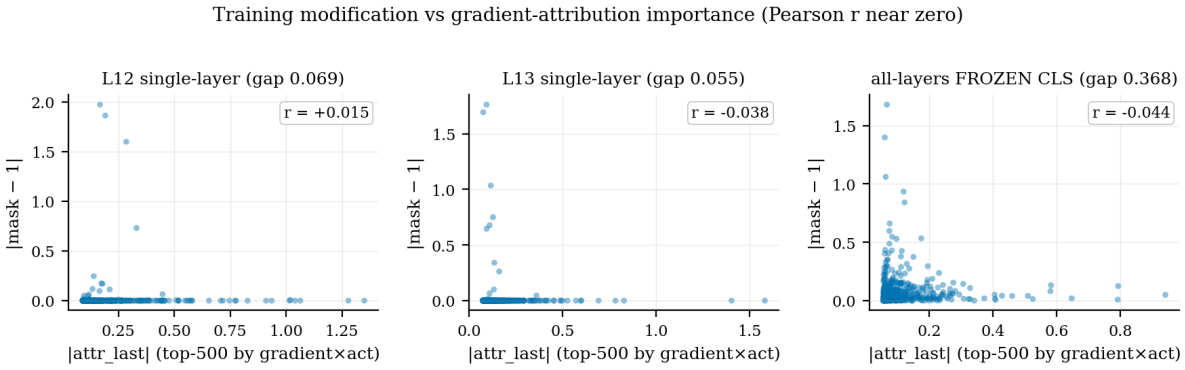

Figure 4: Mask modification magnitude $|M - 1|$ versus gradient-attribution importance |attr_last| for the top-500 neurons by attribution on three Qwen configurations. Pearson correlations are near zero in all cases.

The mask method's interpretability claim is that high-$|M - 1|$ neurons should be the ones the model relies on. We tested this with three independent measurements and found it consistently false. The first is the pairwise top-50 ranking overlap (Figure 3). The four gradient methods cluster together (`attr_last` $\cap$ `attr_sum` ≈ 35, $\cap$ `attr_max` ≈ 43), while `attr_atbad` sits mostly apart from them because it surfaces different circuit components (early-layer token detectors versus late-layer bottleneck integrators). The activation-based `bad_contrast` has moderate overlap with `attr_last` on the shortcut-active FROZEN-CLS checkpoint ($\approx 21/50$), reflecting that gradient methods do pick up genuinely has_bad-responsive neurons when the shortcut is being used. By contrast, `mask_delta` has essentially zero overlap with every other metric on the Qwen all-layers FROZEN-

CLS configuration, including zero overlap with `attr_last`, `attr_atbad`, and `bad_contrast`. The most-modified neurons are not the most-important ones by any measurement. The distribution view in Figure 4 reinforces the same finding: Pearson correlations between $|M - 1|$ and |attr_last| for the top-500 attribution-ranked neurons are near zero across all configurations. The most-attributed neurons cluster at mask values ≈ 1 (training did not modify them, but the model uses them), while the most-mask-modified neurons live in a different part of the rank ordering. The causal ablation results in the next subsection confirm this redundancy directly.

4.4 Ablation: what actually moves the gap

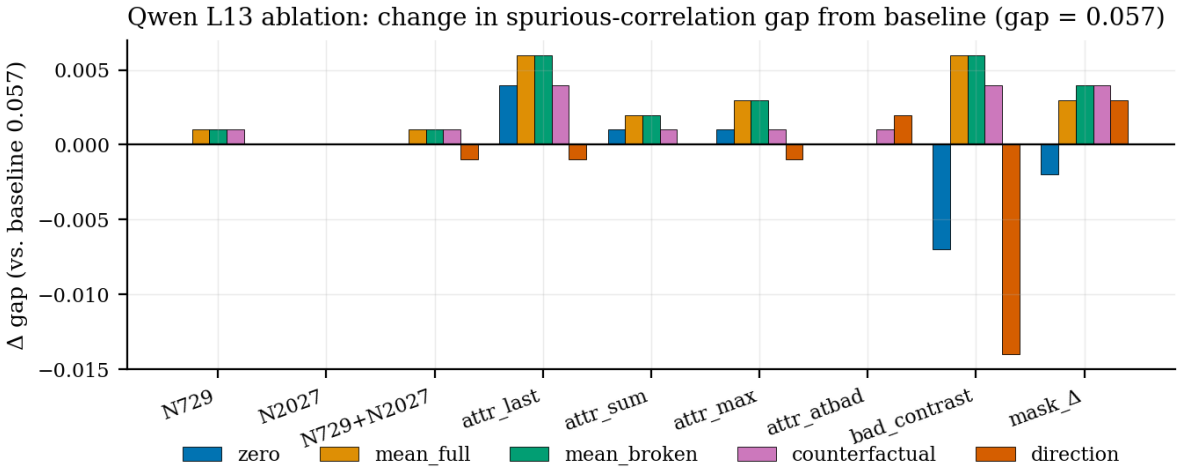

Figure 5: Change in spurious-correlation gap (relative to baseline 0.057) on Qwen L13 across all ablation methods and neuron sets. Most combinations leave the gap unchanged; only `top10_bad_contrast` (under direction or zero ablation) and `top10_attr_max` (under scaling) produce measurable gap reductions.

Table 1: Most-effective ablation per neuron set on Qwen L13 (baseline gap = +0.057, overall accuracy = 0.852).

Neuron Set	Best method	Δ gap	overall	worst
`top10_bad_contrast`	direction	**−0.014**	0.854	0.758
`top10_bad_contrast`	zero	−0.007	**0.856**	0.772
`top10_attr_max`	scaling α=0.5	−0.007	0.853	0.776
`top10_attr_sum`	scaling α=0.5	−0.006	0.853	0.776
`top10_attr_last`	direction	−0.001	0.855	0.772
`top10_attr_atbad`	all five	~ 0	∼base	∼base
`top10_mask_delta`	all five	~ 0	∼base	∼base
`N729`, `N2027`, joint	all five	~ 0	∼base	∼base

Figure 5 and Table 1 summarize the causal ablation results. The cleanest positive result is that direction ablation on the top-10 `bad_contrast` neurons reduces the gap by 24% (0.057 to 0.043), improves overall accuracy from 0.852 to 0.854, and improves the other minority group `neg+no_bad` from 0.880 to 0.906. This is the most effective debiasing intervention we found, and the scaling sweep in Figure 6 confirms it is not a single-point fluke: `top10_bad_contrast` shows a clean monotonic

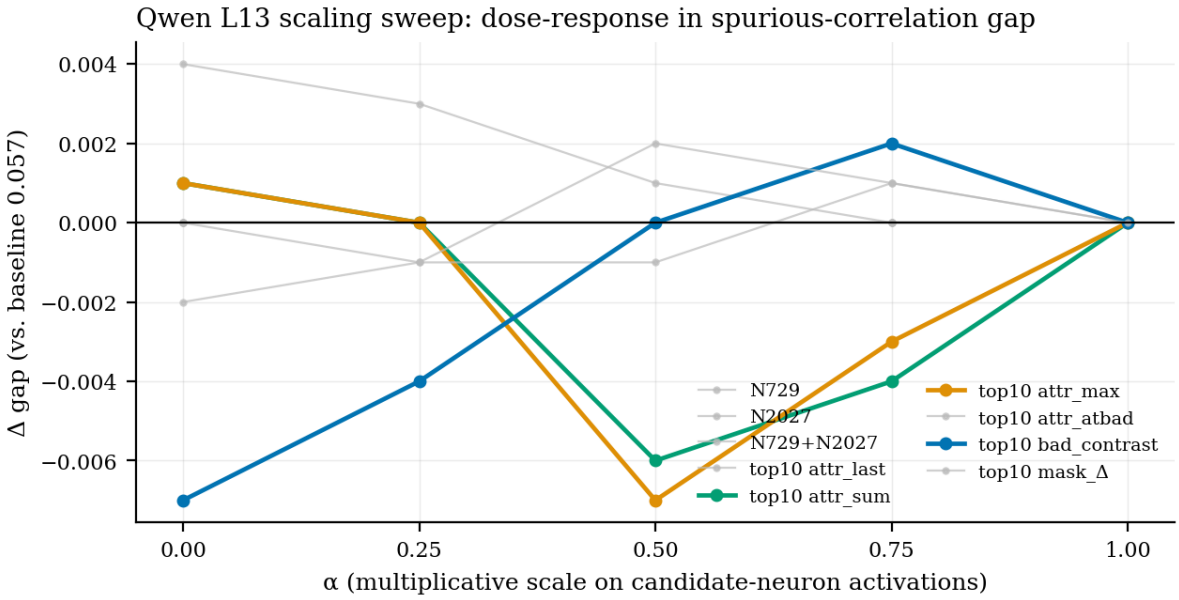

Figure 6: Scaling-sweep dose-response on Qwen L13: change in gap relative to baseline as a function of α (the multiplicative factor applied to candidate-neuron activations; $\alpha = 1$ is no-op). Only `top10_bad_contrast` (and to a lesser extent `top10_attr_max` and `top10_attr_sum`) shows a clean monotonic gap reduction; other sets are flat.

dose-response, with the gap shrinking smoothly as α decreases from 1 to 0. The next-best ranking, `top10_attr_max`, achieves comparable gap reduction at $\alpha = 0.5$, likely because its top-10 includes four `bad_contrast`-top neurons (cross-method overlap). All other neuron sets are essentially flat in the scaling sweep, ruling out cherry-picking and confirming that the causally-implicated neurons are specifically those surfaced by activation contrast.

The negative results are equally clear and form the core of our interpretability finding. Single-neuron ablation is inert: ablating $N729$ or $N2027$ individually or jointly under any of the five methods leaves the metrics within ± 0.002 of baseline, confirming that the spurious computation is distributed rather than carried by any individual neuron. More importantly, the top-10 `mask_delta` set is similarly inert under all five ablation methods, confirming the redundancy finding directly: even the most heavily modified neurons can be set to zero, replaced with their mean, or otherwise disrupted without changing model behavior. The top-10 `attr_atbad` set is also largely inert when ablated at the bottleneck position, consistent with the prior observation that early-layer "bad"-token detectors matter only via their downstream routing through attention rather than via direct bottleneck contribution. Finally, the comparison between ablation methods is methodologically informative. On `top10_bad_contrast`, direction ablation ($\Delta\text{gap} = -0.014$) outperforms zero ablation (-0.007), which itself outperforms mean ablation; mean ablation actually *worsens* the gap by $+0.006$, an out-of-distribution-artifact warning consistent with the intuition that pushing activations toward dataset means can disrupt unrelated computations.

5 Discussion

The capacity-control story works cleanly. The single-layer sweep demonstrates that any single layer's worth of mask capacity is sufficient to learn this task while preventing the catastrophic shortcut,

with the resulting gap bounded tightly in the 0.04–0.08 range and the floor value architecturally invariant (both Qwen and Llama floor at 0.042). The all-layers Qwen explosion to gap 0.44 traces to cross-layer amplification, in which each layer's mask scales up has_bad signal and the head greedily rotates to read it; restricting layer count breaks the feedback loop. This is a practical engineering result: if a practitioner wishes to fine-tune a model that will not overfit a known spurious feature, masking a single layer is sufficient.

The interpretability story does not. The premise of the mask method, namely that $|M - 1|$ identifies the neurons the model relies on, is empirically false on this task on both models. Three independent measurements agree: ranking overlap (0/50 with attribution on Llama all-layers), correlation (Pearson ≈ 0), and causal ablation (zero effect on the gap). The most-modified neurons are not the most-important ones by any criterion. We diagnose this as optimizer degeneracy under co-adapting components. The frozen base provides 117k (Qwen) or 460k (Llama) MLP neurons writing many directions to the residual stream, and the trainable head can rotate to read from whatever the mask leaves available. Many distinct (M, W_{head}) configurations produce equivalent cross-entropy loss because multiple neurons write similar directions, because the head compensates for any mask choice, and because pretrained sentiment features are already strong enough that the mask only needs to nudge them. The optimizer settles into some local-minimum configuration, but that configuration is not uniquely determined by which neurons matter for behavior. This is the classic loss-landscape degeneracy problem: weight changes tell you what the optimizer touched, not what the model uses.

The successful interventions all share a property: they are derived from activation statistics rather than from weight statistics. `bad_contrast` measures activation-distribution shift between has_bad and no_bad inputs and ignores how training got there; direction ablation projects out a direction defined by activation means; counterfactual resampling explicitly intervenes on activations. The gradient-attribution methods sit between, measuring grad-times-act on the fine-tuned model and thus identifying neurons the model uses, though not specifically the ones that encode the spurious feature. The methods that most closely match what we want to find (the spurious-feature carriers) are those that measure what activation pattern responds to the spurious feature.

Several limitations bound these conclusions. The task uses a clean synthetic single-token spurious feature; natural correlations are multi-feature and more entangled. The Llama versus Qwen contrast confounds architecture with a slightly different training-epoch budget (5 versus 10), though mask convergence appears complete in both. The Llama evaluation and attribution JSONs were lost during a directory cleanup mid-project; checkpoints survive and can be regenerated. We deliberately omit Direct Logit Attribution analyses on the LM unembedding W_U because for CLS-head models the LM-unembedding basis is the wrong projection ($\cos(\text{cls\_diff}, W_U[\text{“bad”}]) \approx 0.008$), and instead focus on activation-based diagnostics. Going forward, three concrete experiments would directly probe whether the mask method can be made interpretability-load-bearing. Replacing "bad" with a non-semantic spurious marker would force the mask to construct the shortcut rather than amplify a pretrained one. Switching from a trainable CLS head to a generation-loss objective would remove the head's compensating degrees of freedom. Combining the two, optionally with an explicit training-time penalty on the magnitude of has_bad amplification, would provide the cleanest test of whether sparse-mask training can be steered into producing interpretable updates.

6 Conclusion

We set out to test whether sparse L_0-masked MLP neurons can serve as an inherently interpretable post-training mechanism, and ran a comprehensive empirical study on a synthetic spurious-correlation

task across two model scales. We find a clean separation between two facets of the method. As a capacity-control tool it works as designed, reliably bounding the spurious gap at 0.04–0.08 across all 56 single-layer Qwen and Llama checkpoints and demonstrating that catastrophic shortcut formation requires cross-layer amplification. As an interpretability tool it does not survive causal verification: the modified-mask ranking is uncorrelated with every other importance metric, and ablating the most-modified neurons leaves the gap unchanged. The cleanest debiasing intervention we found uses no mask information at all; it instead identifies neurons by activation contrast between aligned and broken samples and removes their distinguishing direction at the bottleneck. We diagnose the negative result as optimizer degeneracy when an over-parameterized base interacts with a co-adapting readout, and recommend that interpretability claims grounded in weight changes be paired with independent causal verification.

References

[1] Aryaman Arora, Zhengxuan Wu, Jacob Steinhardt, and Sarah Schwettmann. Language model circuits are sparse in the neuron basis. https://transluce.org/neuron-circuits, November 2025.

[2] Tomer Ashuach, Dana Arad, Aaron Mueller, Martin Tutek, and Yonatan Belinkov. Crisp: Persistent concept unlearning via sparse autoencoders, 2025.

[3] Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, Nicholas Joseph, Saurav Kadavath, Jackson Kernion, Tom Conerly, Sheer El-Showk, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Tristan Hume, Scott Johnston, Shauna Kravec, Liane Lovitt, Neel Nanda, Catherine Olsson, Dario Amodei, Tom Brown, Jack Clark, Sam McCandlish, Chris Olah, Ben Mann, and Jared Kaplan. Training a helpful and harmless assistant with reinforcement learning from human feedback, 2022.

[4] Deniz Bayazit, Aaron Mueller, and Antoine Bosselut. Crosscoding through time: Tracking emergence & consolidation of linguistic representations throughout llm pretraining, 2025.

[5] Jan Betley, Niels Warncke, Anna Sztyber-Betley, Daniel Tan, Xuchan Bao, Martín Soto, Megha Srivastava, Nathan Labenz, and Owain Evans. Training large language models on narrow tasks can lead to broad misalignment. *Nature*, 649(8097):584–589, January 2026.

[6] Yiwei Chen, Yuguang Yao, Yihua Zhang, Bingquan Shen, Gaowen Liu, and Sijia Liu. Safety mirage: How spurious correlations undermine vlm safety fine-tuning and can be mitigated by machine unlearning, 2026.

[7] Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. Qlora: Efficient finetuning of quantized llms, 2023.

[8] Melody Y. Guan, Manas Joglekar, Eric Wallace, Saachi Jain, Boaz Barak, Alec Helyar, Rachel Dias, Andrea Vallone, Hongyu Ren, Jason Wei, Hyung Won Chung, Sam Toyer, Johannes Heidecke, Alex Beutel, and Amelia Glaese. Deliberative alignment: Reasoning enables safer language models, 2025.

[9] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin

Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Hanwei Xu, Honghui Ding, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jingchang Chen, Jingyang Yuan, Jinhao Tu, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaichao You, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Mingxu Zhou, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, S. S. Li, Shuang Zhou, Shaoqing Wu, Tao Yun, Tian Pei, Tianyu Sun, T. Wang, Wangding Zeng, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, W. L. Xiao, Wei An, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, X. Q. Li, Xiangyue Jin, Xiaojin Shen, Xiaosha Chen, Xiaowen Sun, Xiaoxiang Wang, Xinnan Song, Xinyi Zhou, Xianzu Wang, Xinxia Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Yang Zhang, Yanhong Xu, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Yu, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yuan Ou, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yunfan Xiong, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Y. X. Zhu, Yanping Huang, Yaohui Li, Yi Zheng, Yuchen Zhu, Yunxian Ma, Ying Tang, Yukun Zha, Yuting Yan, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhicheng Ma, Zhigang Yan, Zhiyu Wu, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, and Zhen Zhang. Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. *Nature*, 645(8081):633–638, September 2025.

[10] Joey Hejna, Rafael Rafailov, Harshit Sikchi, Chelsea Finn, Scott Niekum, W. Bradley Knox, and Dorsa Sadigh. Contrastive preference learning: Learning from human feedback without rl, 2024.

[11] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models, 2021.

[12] Lei Liu, Xiaoyan Yang, Junchi Lei, Yue Shen, Jian Wang, Peng Wei, Zhixuan Chu, Zhan Qin, and Kui Ren. A survey on medical large language models: Technology, application, trustworthiness, and future directions, 2024.

[13] Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in gpt, 2023.

[14] Kevin Meng, Arnab Sen Sharma, Alex Andonian, Yonatan Belinkov, and David Bau. Mass-editing memory in a transformer, 2023.

[15] Aaron Mueller, Atticus Geiger, Sarah Wiegreffe, Dana Arad, Iván Arcuschin, Adam Belfki, Yik Siu Chan, Jaden Fiotto-Kaufman, Tal Haklay, Michael Hanna, Jing Huang, Rohan Gupta, Yaniv Nikankin, Hadas Orgad, Nikhil Prakash, Anja Reusch, Aruna Sankaranarayanan, Shun Shao, Alessandro Stolfo, Martin Tutek, Amir Zur, David Bau, and Yonatan Belinkov. Mib: A mechanistic interpretability benchmark, 2025.

[16] Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback, 2022.

[17] Nikhil Prakash, Tamar Rott Shaham, Tal Haklay, Yonatan Belinkov, and David Bau. Fine-tuning enhances existing mechanisms: A case study on entity tracking, 2024.

[18] Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model, 2024.

[19] Constantin Venhoff, Iván Arcuschin, Philip Torr, Arthur Conmy, and Neel Nanda. Base models know how to reason, thinking models learn when, 2025.

[20] Jason Wei, Maarten Bosma, Vincent Y. Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. Finetuned language models are zero-shot learners, 2022.

[21] Zhengxuan Wu, Aryaman Arora, Zheng Wang, Atticus Geiger, Dan Jurafsky, Christopher D. Manning, and Christopher Potts. Reft: Representation finetuning for language models, 2024.

[22] Yijia Xiao, Edward Sun, Yiqiao Jin, Qifan Wang, and Wei Wang. Proteingpt: Multimodal llm for protein property prediction and structure understanding, 2025.

[23] Yuhang Zhou, Paiheng Xu, Xiaoyu Liu, Bang An, Wei Ai, and Furong Huang. Explore spurious correlations at the concept level in language models for text classification, 2024.

[24] 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, J. Zico Kolter, and Dan Hendrycks. Representation engineering: A top-down approach to ai transparency, 2025.