Introspective Coupling: Self-Explanation Training Tracks Behavioral Change Despite Fixed Supervision

Zifan Carl Guo* **Laura Ruis** **Jacob Andreas** **Belinda Z. Li**
MIT EECS

Abstract

When does training language models (LMs) to generate explanations of their predictions yield faithful introspection, rather than superficial imitation? We study LMs trained to explain which features of their inputs influenced their behavior, using models' counterfactual behavior on modified inputs as supervision. Surprisingly, we find that LMs trained on fixed counterfactual explanations derived from their earlier checkpoints, or even from behaviorally similar models in different families, frequently produce explanations more faithful to *their own current behaviors* than to those of their training targets. This "introspective" coupling between LM explanations and behaviors occurs when training explanations remain sufficiently correlated with current behaviors over the course of training, even as behaviors themselves shift. We also show that introspective coupling tracks behavior shifts: when explanation training is provided concurrently with other post-training objectives, explanations change to match without requiring updated supervision. This phenomenon appears in multiple tasks, including sycophancy and refusal, and is robust to label noise. Overall, our results show that even fixed datasets of counterfactual explanations can provide scalable and generalizable post-training signal for introspection.

1 Introduction

Methods for training language models (LMs) to faithfully output reasons behind their predictions and decisions offer a promising avenue for understanding LM behavior. Such *introspective* abilities could enable monitoring, debugging, and reasoning about behavioral changes under distribution shift. Past work has studied many forms of self-explanation, including explanations of features [35, 21], circuits [26, 29], and behavioral traits [4, 38, 16]. Yet a central challenge is ensuring that LM-generated explanations track an LM's own behavior, rather than merely imitating plausible patterns.

Following past works studying self-explanation in LMs [16, 26, 46], we adopt a counterfactual simulability view of explanation, in which explanations are judged to be faithful if they identify parts of the input (e.g. clues on the answer) that influence model decision-making, and are evaluated by testing whether model behavior changes when (only) identified features are perturbed (Figure 1).

While self-explanation ability may emerge from sufficiently large models [29], out-of-the-box models are not guaranteed to generate faithful explanations [46, 31, 6, 1]. This has motivated a line of work that explicitly trains models to generate explanations. A common recipe is supervised fine-tuning (SFT) on a static set of explanation labels derived from an initial snapshot of the model being explained [4, 28]. This approach has an underexamined limitation: as an LM is trained to generate explanations, its behavior on non-explanation inputs may drift, so the model being trained learns to explain the behavior of the earlier checkpoint, rather than its own, current behavior. However, for a

*Correspondence to carlguo@mit.edu. Code: https://github.com/carlguo866/introspective-coupling.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

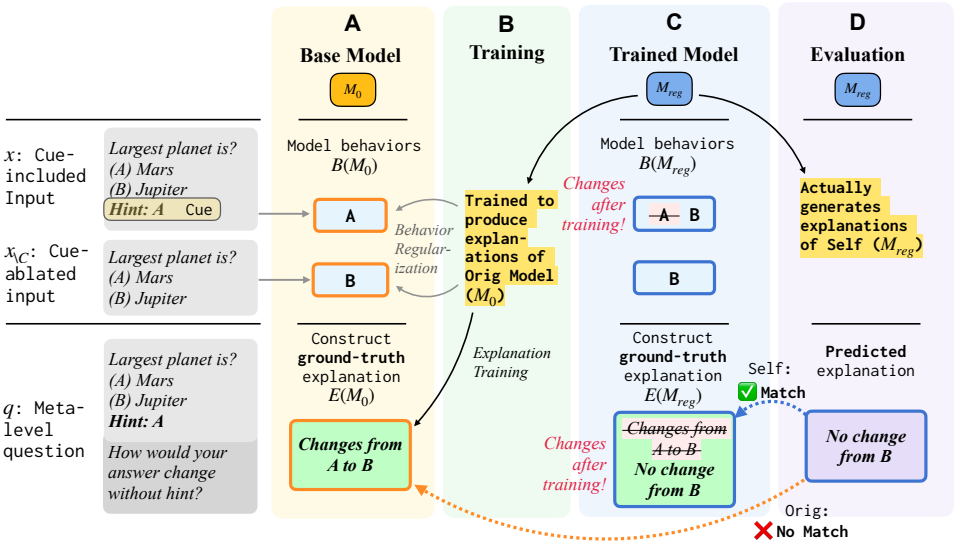


Figure 1: **Method overview.** (A) We sample behaviors $B(\mathcal{M}_0)$ from the base model $\mathcal{M}_0$ on inputs $x, x_{\setminus C}$ and construct explanation labels $E(\mathcal{M}_0)$. (B) We fine-tune $\mathcal{M}_0$ on these explanations, yielding $\mathcal{M}_{\text{reg}}$. (C) During training, $\mathcal{M}_{\text{reg}}$'s behavior drifts to distribution $B(\mathcal{M}_{\text{reg}})$, so its corresponding explanations shift to $E(\mathcal{M}_{\text{reg}})$. (D) $\mathcal{M}_{\text{reg}}$'s output explanations track this change: they better match explanations of $\mathcal{M}_{\text{reg}}$'s own behavior (Self) despite training on explanations of $\mathcal{M}_0$ (Orig). We call this Self $>$ Orig *introspective coupling*.

self-explanation to be faithful, it must track the model's current behavior rather than a stale set of behavior before explanation training.

In this paper, we make the surprising observation that training models to generate explanations of their earlier behavior, while regularizing behaviors themselves toward an earlier checkpoint, *in fact causes them to explain their current behavior more faithfully than the checkpoint from which the explanations were derived* (Figure 1). On test data, regularized explanation training reduces drift relative to unregularized fine-tuning, but does not eliminate it entirely. A trained model's explanations track this residual drift, despite never receiving supervision derived from its current, on-policy behavior. This effect disappears without regularization.

More formally, we call this effect *Self > Orig*, where:

> A model demonstrates **Self > Orig** if its explanations more faithfully predict its current behavior than the original behavior used to generate its training labels.

We interpret Self > Orig as evidence of *introspective coupling*:

> **Introspective Coupling**: a phenomenon where a model learns to couple its explanations to its own current behavior rather than merely reproducing static explanation training targets.

We show that **Self > Orig** is robust across three tasks (Section 3): two sycophancy datasets (Hint-MMLU [6] and AITA [7]), and a refusal dataset comprising a mix of FalseReject [51] and WildJailbreak [19]. In each case, regularized training produces models whose explanations are more faithful to their *current* behavior than to the behavior of the original model that generated their training labels.

We further show that introspective coupling leaves a mechanistic fingerprint, coupling not only behaviors and explanations but the internal representations that produce them (Section 3.3).

We characterize *when* introspective coupling emerges (Section 4), finding that explanation training data must remain similar to a model's *current, online* behavior—even if explanations diverge from the model's *original* behavior by up to 50%. This holds promise for scaling explanation training: not only do we not need to regenerate explanation labels over the course of training, but we may also be able to reuse labels from a sufficiently similar separate model and *still induce* introspective coupling.

Finally, we demonstrate that introspective coupling remains useful when a model acquires new behavior beyond the original explanation labels (Section 5). We train models on auxiliary data that alters their underlying behavior distribution, either directly or indirectly, and find that explanation

tracks those changes. This is promising for integrating introspection training into standard post-training pipelines: models' explanations can track behavioral drifts induced by different supervisions.[2]

2 Methods

2.1 (Counterfactual) Explanation Construction

Following past work, we focus on counterfactual explanations: how behaviors change under an edit to its input (e.g. removal of a hint) [26, 16]. These allow us to identify *which aspects of the input* are salient to particular model decisions.

Let a **cue** $C \subseteq x$ denote a contiguous span of an input x hypothesized to causally influence a model's behavior. We write $x_{\setminus C}$ for the **cue-ablated input** obtained by removing C from x, and refer to the original x as the **cue-included input**. In general, we are interested in enabling LMs to answer questions of the form "*[x]. If the cue were removed, how would the assistant's answer change?*" with answers of the form "*The response [would/would not] change to* <$\mathcal{M}(x_{\setminus C})$>." We train models to produce these answers via supervised fine-tuning (SFT). We use notation shown in Table 1 throughout. As illustrated in Figure 1, our method has four steps:

1. $\mathcal{M}_0$ **explanation construction (§2.1):** We start with base model $\mathcal{M}_0$, sample model behaviors $B(\mathcal{M}_0)$, from which ground-truth explanations $E(\mathcal{M}_0)$ are constructed. In the running example of Figure 1, $\mathcal{M}_0$ answers A with the hint and B without it, so $E(\mathcal{M}_0)$ is "behavior changes from A to B when the hint is removed".
2. **Explanation training (§2.2):** We train $\mathcal{M}_0$ on explanations $E(\mathcal{M}_0)$ while regularizing its behaviors towards $B(\mathcal{M}_0)$, obtaining a model:
$$\mathcal{M}_{\text{reg}} = \mathcal{M}[B(\mathcal{M}_0), E(\mathcal{M}_0)]. \tag{1}$$
3. $\mathcal{M}_{\text{reg}}$ **explanation construction (§2.1):** We roll out behaviors $B(\mathcal{M}_{\text{reg}})$ from the new $\mathcal{M}_{\text{reg}}$ and construct ground-truth explanations $E(\mathcal{M}_{\text{reg}})$ describing those behaviors. In Figure 1, $\mathcal{M}_{\text{reg}}$ answers B both with and without the hint, so $E(\mathcal{M}_{\text{reg}})$ is "behavior remains unchanged as B."
4. **Measuring introspective coupling (§2.3):** We evaluate whether the predicted explanations generated by $\mathcal{M}_{\text{reg}}$ better match the ground-truth explanations of the original model $E(\mathcal{M}_0)$, or the explanations of self $E(\mathcal{M}_{\text{reg}})$. As shorthand, we denote evaluating $\mathcal{M}_{\text{reg}}$'s predictions on $\mathcal{M}_0$'s explanations as $\mathcal{M}_{\text{reg}} \to \mathcal{M}_0$ (or Orig) and evaluating $\mathcal{M}_{\text{reg}}$'s predictions on $\mathcal{M}_{\text{reg}}$'s explanations as $\mathcal{M}_{\text{reg}} \to \mathcal{M}_{\text{reg}}$ (or Self). **When Self > Orig, we say that a model exhibits introspective coupling.** In Figure 1: $\mathcal{M}_{\text{reg}}$'s output matches $E(\mathcal{M}_{\text{reg}})$ but not $E(\mathcal{M}_0)$, and thus, introspective coupling is observed.

2.2 Explanation Training

Given a model $\mathcal{M}$, we perform SFT training on explanation data from base model $E(\mathcal{M}_0)$ while regularizing behaviors to be close to $B(\mathcal{M}_0)$. We use cross-entropy to train explanations and KL divergence to regularize behaviors. Formally, the objective is:

$$\mathcal{L}(\mathcal{M}) = \underbrace{\mathbb{E}_{(q,e)\sim E(\mathcal{M}_0)}\big[-\log p_{\mathcal{M}}(e \mid q)\big]}_{\text{explanation cross-entropy}} + \lambda \underbrace{\mathbb{E}_{(x,y)\sim B(\mathcal{M}_0)}\big[\mathrm{KL}(p_{\mathcal{M}_0}(y \mid x) \,\|\, p_{\mathcal{M}}(y \mid x))\big]}_{\text{behavioral regularizer}} \tag{2}$$

2.3 Evaluation Metrics

We evaluate all models on two main metrics: **(1) Explanation Match (EM)** scores how well the LM $\mathcal{M}$ has learned to explain a target model $\mathcal{M}'$, where $\mathcal{M}' = \mathcal{M}_0$ for Orig and $\mathcal{M}' = \mathcal{M}_{\text{reg}}$ for Self. Specifically, we measure exact match between the explainer $\mathcal{M}$'s prediction $\hat{e}^{(\mathcal{M})} \sim \mathcal{M}(\cdot \mid q)$ for each meta-question q ("if the cue were removed, how would your answer change?") and the ground-truth explanation in $E(\mathcal{M}')$, averaged over $E(\mathcal{M}')$.

[2]We refer to "introspection" in the context of the operationalization above where models exhibit **behavioral** introspection that matches its own object-level outputs. We emphasize that there is an ongoing debate about the specific definition of model introspection and do not make claims that models have metacognitive awareness or complete access to all internal computations.

(2) Behavioral Match (BM) captures the drift between two models' behavioral outputs, such as the drift of a model before and after explanation training. Let $\mathcal{D} = \{(x, C)\}$ be a held-out set of inputs x and cue spans C, and let $y_{\mathcal{M}}(x) \sim \mathcal{M}(\cdot \mid x)$ denote $\mathcal{M}$'s label on x, and $y_{\mathcal{M}}(x_{\setminus C})$ denote the label on $x_{\setminus C}$. We measure agreement between $\mathcal{M}$ and $\mathcal{M}_0$ on both versions of each input:

$$\mathrm{BM}(\mathcal{M}, \mathcal{M}_0) \;=\; \frac{1}{|\mathcal{D}|} \sum_{(x,C)\in\mathcal{D}} \underbrace{\mathbf{1}\big[y_{\mathcal{M}}(x) = y_{\mathcal{M}_0}(x)\big]}_{\text{cue-included match}} \cdot \underbrace{\mathbf{1}\big[y_{\mathcal{M}}(x_{\setminus C}) = y_{\mathcal{M}_0}(x_{\setminus C})\big]}_{\text{cue-ablated match}}. \tag{3}$$

In the metric above, both sides of the behavior must match. Fine-grained metrics are in Appendix E.1. We report error bars as 95% bootstrap CIs and significance with exact McNemar tests.

3 Characterizing Introspective Coupling

3.1 Experiment Setup

Models. Our primary model $\mathcal{M}_0$ is **Qwen3-8B** [50]. To verify that our result is not model-specific, we show results across model families and sizes in Appendix E.7. We use full fine-tuning by default and discuss LoRA fine-tuning in Appendix G.1.

Sycophancy Datasets. Modern language models often exhibit sycophancy—over-agreeableness with the user—at the expense of accuracy [40, 49]. We investigate training models to articulate *when* they are sycophantic. Specifically, we study two datasets: First, **Hint-MMLU** [6, 26] assesses whether models will modify their answer to follow a user-suggested hint. Here, x is an MMLU multiple-choice question with an injected $C =$"Hint: `A`" string (see Figure 1). Sycophantic models will change their answer when the hint is present; we train models to articulate whether they will do so. Second, **AITA** measures whether models are overly inclined to flatter the user [5]. This dataset consists of morally ambiguous stories from Reddit's *r/AmItheAsshole*. Here, x is a Reddit post and y is a judgment of the user's moral character $\in$ {"Not the Asshole (NTA)", "You're the Asshole (YTA)"}. C is a system prompt framing the story in first- or third-person. In this case, sycophantic models will change their judgment to be more favorable if the story were written in first-person. We train models to articulate when they will change their judgment based on the system prompt.

Refusal Datasets. Understanding when LMs refuse requests allows us to debug over- or under-refusals. We specifically investigate how the user's presented *role* interacts with refusal. The user's role is a popular attack surface—models may be more susceptible to answering harmful queries for users who are fiction writers, or to overrefusing users who are kindergarten teachers. Thus, we would like to train models to articulate whether and how their refusal hinges on a stated role. x is a user prompt and C is a system prompt that frames the user in different roles. See examples in Appendix E.6. We measure $y \in \{\text{refuse}, \text{comply}\}$ by sampling a freeform response and scoring it with an LLM judge. Note that this setting is much harder than Hint-MMLU and AITA: the behavior label is inferred from freeform generation rather than constructed deterministically from a single answer token, so the explanation target is a higher-level property of the response. We obtain x by combining two refusal datasets: **FalseReject** [51], a benign dataset to test over-refusal, and **WildJailbreak** [19] for harmful queries. We generate roleplay system prompts C with an LLM.

3.2 Results

Results can be found in Figure 2, where we compare regularized training ($\lambda = 1$ in Equation (2), left side) against unregularized training ($\lambda = 0$ in Equation (2), right side), plotting both behavior drift (a,c) and Explanation Match for both Self and Orig (b,d).

From Figure 2(b), we observe that across held-out data on all three tasks—Hint-MMLU, AITA, and Refusal—the regularized model $\mathcal{M}_{\mathrm{reg}}$'s predicted explanations match its own labels $E(\mathcal{M}_{\mathrm{reg}})$ better than the original targets $E(\mathcal{M}_0)$ it was trained on. Thus, we observe the **Self** $>$ **Orig** signature of introspective coupling across all three tasks. By construction, this gap arises entirely from examples where the two labels disagree. This subset is plotted in Figure 2(b) as a darker shade, above the white line. On this subset, $\mathcal{M}_{\mathrm{reg}}$ resolves the discrepancy in favor of its own behavior 79%, 82%, and 63% of the time, respectively. In Appendix E.2, we rule out trivial accounts explaining **Self** $>$ **Orig**: that $E(\mathcal{M}_{\mathrm{reg}})$ drifts to a degenerate distribution that is easier for *any* external explainer to learn. We validate that regularized training produces this signature across model families and sizes (Figure 15).

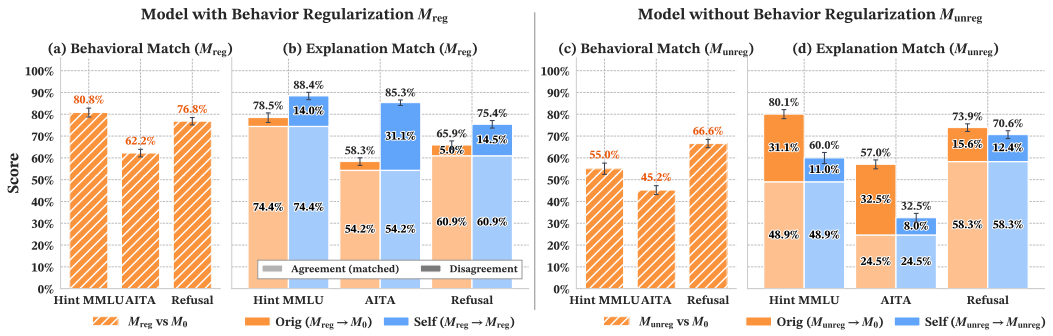


Figure 2: Self > Orig emerges only with regularization, across Hint-MMLU, AITA, and Refusal domains. (**a**, **b**): the regularized model $\mathcal{M}_{\text{reg}}$. (**c**, **d**): the unregularized model $\mathcal{M}_{\text{unreg}}$. **(a, c)** Behavioral Match: agreement between original labels $B(\mathcal{M}_0)$ and labels after training ($B(\mathcal{M}_{\text{reg}})$ and $B(\mathcal{M}_{\text{unreg}})$ respectively). **(b, d)** Explanation Match: explanations against $E(\mathcal{M}_0)$ labels (orange) and self labels (blue). $\mathcal{M}_{\text{reg}}$, despite being trained on $\mathcal{M}_0$, explains itself better than $\mathcal{M}_0$ (b); without regularization, this gap reverses (d). Much of $\mathcal{M}_{\text{reg}}$'s apparent ability to explain $\mathcal{M}_0$ comes from the subset where $\mathcal{M}_0$ and $\mathcal{M}_{\text{reg}}$ *agree* (the lighter lower segment in panel b), and the gap comes from the disagreement subset.

From Figure 2(d), we observe that **Self > Orig** disappears without regularization: a model trained on explanations alone, $\mathcal{M}_{\text{unreg}} = \mathcal{M}[E(\mathcal{M}_0)]$, actually demonstrates Self < Orig. When comparing Figure 2(a) against Figure 2(c), we notice that $\mathcal{M}_{\text{unreg}}$'s behavior drifts further away from $\mathcal{M}_0$ than $\mathcal{M}_{\text{reg}}$'s behavior drifts. Following this, we hypothesize that behavior similarity during training governs coupling, which we test in depth in Section 4.

3.3 Introspective coupling at the mechanistic level

One hypothesis for why coupling emerges is that through training, computation that *reports* the model's behavior learns to reuse the computation that *produces* the behavior. We investigate whether the two representations are causally linked [33] at the mechanistic level, by testing if intervening on representations to shift behavior $B(\mathcal{M})$ also shifts the explanation $E(\mathcal{M})$ in the same direction.

We test this on Hint-MMLU models via activation patching (Figure 3 left). Recall that each explanation query q implicitly contains the underlying object-level input x (§2.1). We patch activations into the shared prefix x and measure the downstream effect on both behaviors $\mathcal{M}(x)$ and explanations $\mathcal{M}(q)$. We create two counterfactual inputs, a source x and a target x' (with corresponding meta-questions q, q') that share the same hint cue $C = C'$, but differ in behavior, $\mathcal{M}(x) \neq \mathcal{M}(x')$, and hence in explanation, $E(\mathcal{M})(q) \neq E(\mathcal{M})(q')$. We patch activations from x to x' (and from q into q') at corresponding (layer, token) positions, and measure the normalized logit difference (NLD) [48]

$$\text{NLD}(u) = \frac{\ell_{\text{patched}} - \ell_{u'}}{\ell_u - \ell_{u'}}, \qquad u \in \{x, q\}, \tag{4}$$

where ℓ is the output logit of each forward pass. The NLD is 0 if the patch has no effect and 1 if it fully recovers the source output.

We validate the following hypothesis: if the model shares circuits between behavior and explanation, then patching interventions that affect behavioral logits $\text{NLD}(x)$ are correlated with ones that affect explanation logits $\text{NLD}(q)$. The result is shown in Figure 3, where interventions move the behavior logits and explanation logits in the same direction, with Pearson correlation $r = 0.89$. This means that the same layer/token activations are used for behavioral responses as for the explanations. Critically, the $\mathcal{M}_{\text{unreg}}$ baseline, which has non-trivial Orig Explanation Match (Figure 2(d)) but does not exhibit **Self > Orig**, only exhibits a correlation of $r = 0.527$ (Appendix F).

4 When does introspective coupling happen?

Introspective coupling is desirable for faithful explanation because it means that the model's explanations track its current behavior. We next ask **when this coupling emerges.** Our findings in Section 3.2 suggested that models only demonstrated Self > Orig under regularization, which reduced behavioral

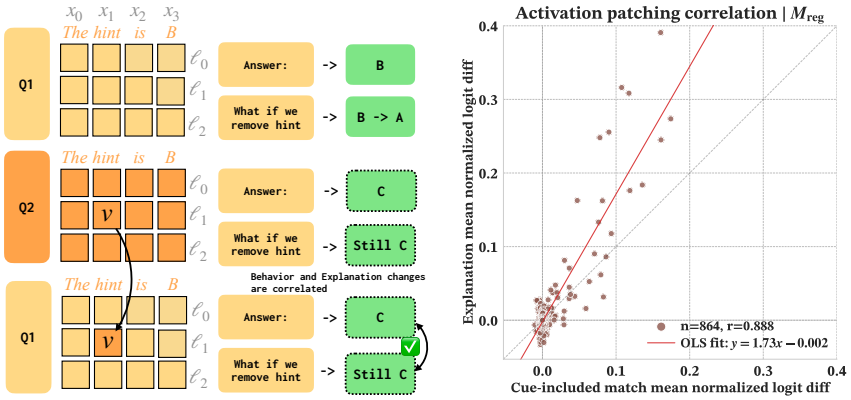


Figure 3: Mechanistic signature of introspection: activation-level interventions that change behavior also change the explanation. **Left:** we construct prompt pairs with the same cue but different explanation labels (removing the cue changes the answer for one, but not the other). We patch activations between them and measure whether shifts in behavior logits correlate with shifts in explanation logits. **Right:** Each point is one (layer, token) patching position. X-axis is the mean normalized logit difference for behavior (cue-included answer), while the y-axis shows the corresponding logit-difference for the explanation's change/no-change prediction. Patching effects on behavior and explanation are strongly correlated with Pearson $r = 0.89$.

drift compared to no regularization. Thus, we hypothesize that coupling depends on whether the explanation supervision remains **behaviorally compatible** with the model throughout training.

Formally, let $E^{(t)}_{\text{sup}}$ denote the explanation supervision labels used for SFT at training step t. In the default fixed-label setting, $E^{(t)}_{\text{sup}} = E(\mathcal{M}_0)$ for all t; in this section, we vary $E^{(t)}_{\text{sup}}$ directly. Let $E(\mathcal{M}_t)$ denote the ground-truth explanations constructed from the current model's behavior at step t. We refer to the similarity between $E^{(t)}_{\text{sup}}$ and $E(\mathcal{M}_t)$ as **online label–self agreement**: the degree to which the explanation training targets at step t agree with the explanations of the current model's behavior. Thus, our central hypothesis is

> **Hypothesis:** High online label-self agreement (i.e. $E^{(t)}_{\text{sup}}$ and $E(\mathcal{M}_t)$ agreement) over the course of training (i.e. $\forall t$) is an important factor for the emergence of Self $>$ Orig.

The three subsections below probe this hypothesis by varying online agreement through different mechanisms: changing the behavioral regularization weight λ, which affects how close $\mathcal{M}_t$ remains to the fixed explanation labels (Section 4.1); directly controlling the agreement between $E^{(t)}_{\text{sup}}$ and $E(\mathcal{M}_t)$ via online relabeling (Section 4.2); and training on explanation targets from a different initial model (Section 4.3), which separates the role of *initial* label–model similarity from *online* compatibility during training.[3] Finally, we also perform a learning rate sweep that uncovers evidence *against* this hypothesis, suggesting that multiple factors may be necessary to induce Self > Orig (Appendix G.2). We focus our analysis in this section on the Hint-MMLU setting.

4.1 Behavioral Regularization Preserves Online Label–Self Agreement

We begin by performing a finer-grained version of our regularization vs. no regularization experiment in Section 3.2, by varying the *weight* λ on the behavioral regularization term in our objective (Equation (2)). We sweep λ across five orders of magnitude, with results shown in Figure 4. We find that coupling begins to emerge at remarkably small values: a Self $>$ Orig gap is visible for any $\lambda \gtrsim 5\times10^{-3}$, and the gap stays consistent for bigger λ. Notably, this inflection coincides with a large increase in Behavioral Match (Figure 4a) between $\mathcal{M}_t$ and $\mathcal{M}_0$. This is consistent with our hypothesis that high online label-self agreement is important for Self > Orig. In this setup, the original model $\mathcal{M}_0$ is the supervision source throughout training so $E^{(t)}_{\text{sup}} = E(\mathcal{M}_0)$ for all t. Thus, when $\mathcal{M}_t$ and $\mathcal{M}_0$ remain close (high Behavioral Match) throughout training, $E(\mathcal{M}_t)$ remains similar to $E^{(t)}_{\text{sup}}$, and Self > Orig emerges. However, this experiment does not distinguish whether coupling requires the current model being trained to remain close to the original model $\mathcal{M}_0$ or to the source

[3]We also plot the emergence of coupling over the course of training in Appendix D.

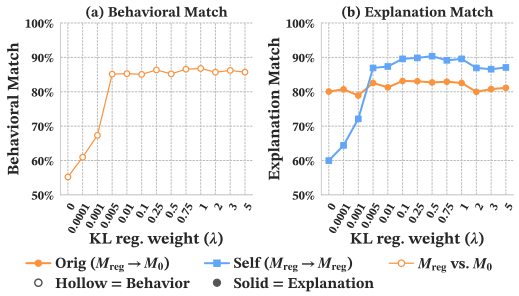


Figure 4: Behavioral regularization weight λ sweep on Hint-MMLU (§4.1). **(a)** Behavioral Match between $\mathcal{M}_{\text{reg}}$ and $\mathcal{M}_0$ rises at the same λ-values as Self Explanation Match shown in **(b)**, which shows the Self > Orig signature and overtakes Orig once $\lambda \geq 5\text{e}{-3}$ and stays ahead for bigger λ.

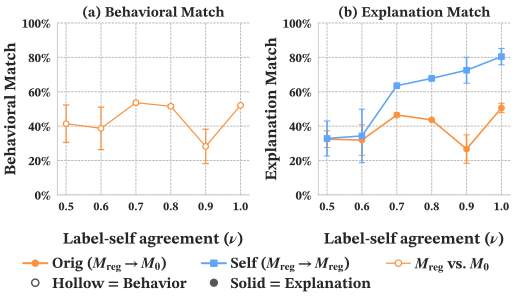


Figure 5: Continuous relabeling with adjustable label–self agreement ν between training targets $E_{\text{sup}}^{(t)}$ and self-labels $E(\mathcal{M}_t)$ (§4.2). **(a)** Behavioral Match stays constant at ≈ 0.5, so Self > Orig does not come from reduced drift from $\mathcal{M}_0$. **(b)** Meanwhile, Self Explanation Match rises sharply over Orig once $\nu \approx 0.7$.

of the explanation labels $E_{\text{sup}}^{(t)}$. We isolate this distinction directly in Section 4.2, where we vary the current model's proximity to the labels, independent of the original model.

4.2 Introspective Coupling Depends on Online Label-Self Agreement

We conduct a controlled test of our hypothesis that high online label-self agreement leads to introspective coupling. Because $\mathcal{M}$'s behavior changes dramatically over the course of training, at every gradient step, we regenerate the explanation training targets $E_{\text{sup}}^{(t)}$ so that they agree with the *current* model behavior $B(\mathcal{M})$ on a fraction ν of labels and disagree—via a randomly sampled alternative answer—on the remaining fraction $1-\nu$:

$$E_{\text{sup}}^{(t)} = \left\{ (q, e) \;\middle|\; e = \begin{cases} B(\mathcal{M})(q) & \text{with probability } \nu \\ \tilde{e} \neq B(\mathcal{M})(q) & \text{with probability } 1-\nu \end{cases} \right\}$$

On the behavioral side, we still regularize $\mathcal{M}$ toward the base model $\mathcal{M}_0$. Thus, we have a model:

$$\mathcal{M}_{\text{noise-self}} = \mathcal{M}[E_{\text{sup}}^{(t)}, B(\mathcal{M}_0)].$$

We sweep ν in Figure 5. In (b), we find a sharp transition in explanation behavior as the online label–self agreement ν varies. We observe introspective coupling (Self > Orig) for $\nu \geq 0.7$ but once $\nu \leq 0.6$, both Self and Orig Explanation Match collapse. This supports our hypothesis that coupling requires a minimum level of agreement between the explanation supervision and the current model's behavior; for Qwen3-8B on Hint-MMLU, this threshold also appears to be ≈ 0.7.

We measure Behavioral Match between the trained model and the original model $\mathcal{M}_0$ in Figure 5(a). Interestingly, Behavioral Match remains roughly constant across ν. This disentangles the two factors we suggested at the end of the previous section: staying close to the *explanation labels* during training matters more than staying close to the *original model*.

4.3 Introspective Coupling Persists with Cross-Model Explanation Labels

Previously, we found that it is more important for the current model to remain close to the explanation labels than to the initial model. We next ask whether those labels must be generated by the model's own initial checkpoint. **Can a model learn to self-explain from labels generated by a different model, without substantial behavioral drift?** If so, explanation labels need not be collected separately for every model: labels from one model could induce self-explanations in another.

To test this, we start with a Qwen3-8B base model $\mathcal{M}_0$, from which we derive behaviors $B(\mathcal{M}_0)$ and ground-truth explanations $E(\mathcal{M}_0)$. We then replace some of Qwen's explanation labels with labels from Llama-3.1-8B-Instruct, denoted $\mathcal{M}_{\text{Llama}}$. The two models' ground-truth explanations agree on only 53% of examples.

For each mixture ratio $\alpha \in [0, 1]$, we construct an explanation supervision dataset

$$E_\alpha = (1-\alpha)\, E(\mathcal{M}_0) + \alpha\, E(\mathcal{M}_{\text{Llama}}),$$

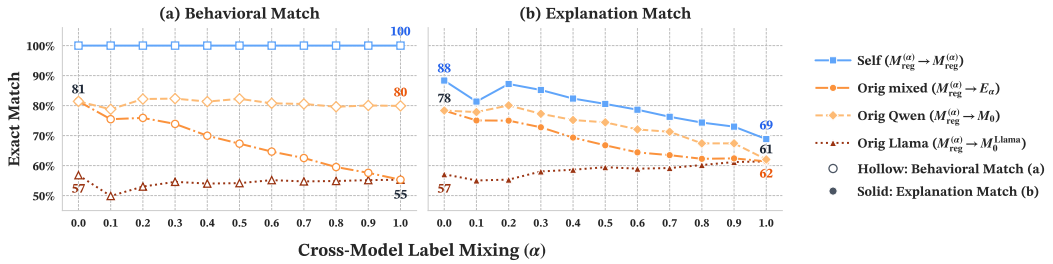


Figure 6: Training on explanation labels from another model (§4.3). Explanation labels E_α mix a fraction of Llama-3.1-8B-Instruct labels into Qwen3-8B's labels ($\alpha = 0$: all Qwen's labels, $\alpha = 1$: all Llama's). Behavior is still regularized toward Qwen. We score each trained model against four label sources: Self, E_α (mixed), Qwen-base, and Llama-base. **(a)** Across the sweep, behavior remains close to initial Qwen (≈80% Behavioral Match), and **(b)** Self > Orig holds at every α. Though explanation supervision comes from another model, the trained model's explanations continue to track its own current behavior better than either base-model label source.

where α is the fraction of explanation labels sourced from Llama. We train

$$\mathcal{M}_{\text{mix}} \;=\; \mathcal{M}[B(\mathcal{M}_0),\; E_\alpha]\,.$$

Crucially, we regularize behavior toward the original Qwen throughout, using $B(\mathcal{M}_0)$; only the explanation supervision is mixed. Thus, $\alpha = 0$ corresponds to pure Qwen explanation supervision, while $\alpha = 1$ corresponds to pure Llama explanation supervision with Qwen behavioral regularization.

Results are shown in Figure 6. Panel (a) confirms that this intervention does not significantly affect the trained model's behavior. Across α, $\mathcal{M}_{\text{reg}}^{(\alpha)}$ remains a steady 80% from $\mathcal{M}_0$ (base Qwen) and a steady 55% from $\mathcal{M}_{\text{Llama}}$ (base Llama). Thus, replacing Qwen explanation labels with Llama labels does not cause the trained model's behavior to move toward Llama. Nevertheless, on the explanation side (b), Self $>$ Orig holds for every value of α. Thus, **introspective coupling does not require explanation labels to come from the same model being trained**.[4]

These results suggest that explanation datasets may be reusable across behaviorally similar models. Further, these results add to a growing body of work arguing that self-explanation in LMs leverages *privileged access* [4, 26]. Even when the model is explicitly trained on another model's explanations, including examples where the two models disagree, its predictions track its own behavior more closely than the foreign explanation labels. We interpret this as evidence that introspective coupling depends more on *online label–self agreement* than on the source of the original explanation labels.

5 Generalization

In Sections 3 and 4, we studied behavioral drift that occurs as an emergent consequence of explanation training. However, in realistic post-training scenarios (or in scenarios where explanations would be useful), models are trained concurrently with other post-training data that may induce different behavioral shifts. **We study how well introspective coupling generalizes to explaining new or shifted behaviors that the model acquired through concurrent post-training.**[5] To formalize this setup, we extend our notation to include a dataset argument: let $B(\mathcal{M}, \mathcal{D})$ denote behaviors produced by model $\mathcal{M}$ on dataset $\mathcal{D}$ (e.g. Hint-MMLU), and let $E(\mathcal{M}, \mathcal{D})$ denote explanations corresponding to those behaviors. Starting from our default training pipeline, we concurrently train $\mathcal{M}$ on an auxiliary post-training dataset A:

$$\mathcal{M}_{\text{aux}} \;=\; \mathcal{M}[B(\mathcal{M}_0, \mathcal{D}),\; E(\mathcal{M}_0, \mathcal{D}),\; A] \tag{5}$$

where explanation training and behavioral regularization are applied only on the original dataset $\mathcal{D}$ and the auxiliary dataset A provides *behavior-only* supervision, not explanation supervision.

[4] That said, explanation quality does degrade overall as α increases: Self Explanation Match falls from 88% ($\alpha = 0.0$) to 69% ($\alpha = 1.0$), as expected if Llama labels act as noise to the function of modeling Qwen's own behaviors.

[5] A version of this question was explored by Binder et al. [4] using synthetic tasks; we focus on more complex explanations with direct safety relevance, such as sycophancy or refusal behavior, under realistic post-training mixtures.

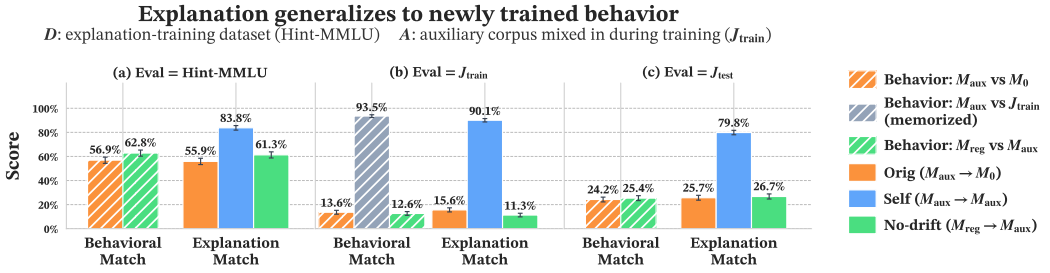


Figure 7: Explanations generalize to newly acquired behaviors (§5.1). We train $\mathcal{M}_{\text{aux}}$ with explanation supervision on Hint-MMLU (D) plus behavior-only training on gibberish data J_{train} (A), and evaluate on held-out Hint-MMLU, J_{train}, and J_{test}. Self (blue) > Orig (orange) holds on all three sets. Although it never saw J_{train} *explanations* during training, nor J_{test} explanations or behavior, $\mathcal{M}_{\text{aux}}$ explains both at high accuracy. $\mathcal{M}_{\text{reg}}$, which never saw Jabberwocky, can explain $\mathcal{M}_{\text{aux}}$ only at chance.

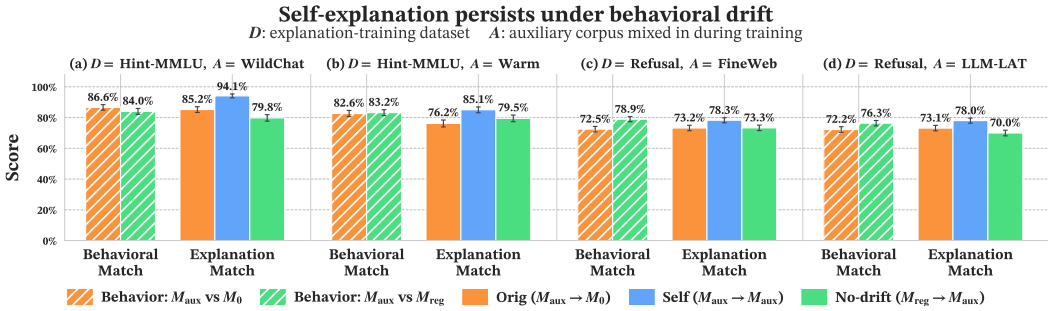


Figure 8: Explanations track shifts to existing behaviors (§5.2). We mix auxiliary post-training corpora into explanation training (Equation (5)): each panel **(a)–(d)** mixes an explanation-training dataset $\mathcal{D}$ with an auxiliary corpus A. **Behavioral**: every setup drifts from both $\mathcal{M}_0$ and $\mathcal{M}_{\text{reg}}$ (explanation training only). **Explanation**: We find that (1) Self > Orig persists under auxiliary training, and (2) Self > $\mathcal{M}_{\text{reg}} \rightarrow \mathcal{M}_{\text{aux}}$, showing that the model's explanations have shifted alongside auxiliary training to reflect its shifted behavior.

We then evaluate whether $\mathcal{M}_{\text{aux}}$'s explanations track its current behavior on $\mathcal{D}$ and on A. Ground-truth explanations are constructed post hoc from $\mathcal{M}_{\text{aux}}$'s own behavior. Thus, faithful explanations on A cannot result from direct supervision and must reflect generalization of the explanation function. We study two kinds of auxiliary shift: in §5.1 we inject new, synthetic behavior, and in §5.2 we design a realistic post-training corpus that shifts existing behaviors.

5.1 Generalization to Newly Acquired Behaviors

We first test whether coupling extends to behaviors the model could only have acquired through auxiliary training. We generate synthetic nonsense "Jabberwocky" data $J = \{(x_J, y_J)\}$ that follows the Hint-MMLU question format but uses fictitious, alien-sounding questions that LMs do not have priors over. The model's behavior on J must be trained in (samples in Appendix H.1). We use the objective in Equation (5), where we train on explanations from $D =$ Hint-MMLU and on the Jabberwocky auxiliary dataset $A = J_{\text{train}}$, holding out J_{test} to test generalization.

Results. Figure 7 reports Explanation Match on the three evaluation sets ($\mathcal{D}$, A, J_{test}). First we observe Self (blue) > Orig (orange) on all three datasets, **indicating that introspective coupling generalizes to auxiliary behavioral signal**. In particular, (a) shows that the coupling on the original dataset $\mathcal{D}$ is preserved under auxiliary signal, (b) shows that $\mathcal{M}_{\text{aux}}$ explains J_{train} nearly perfectly (90.5%) despite never receiving explanation supervision, and (c) shows that explanation accuracy is 79.8% on J_{test}, which received neither explanation nor behavior signal.[6] These results are not explained by the auxiliary data failing to change the model: $\mathcal{M}_{\text{reg}}$, which never sees Jabberwocky data, explains $\mathcal{M}_{\text{aux}}$ only at chance. Thus, $\mathcal{M}_{\text{aux}}$'s explanations track behaviors acquired through additional training rather than merely preserving its original explanation policy.

[6] J_{test} responses are near random and have no ground-truth, and the model defers to changing to the hint. However, $\mathcal{M}_{\text{aux}}$ knows to output the *cue-ablated* random response, an ability that $\mathcal{M}_0$ or $\mathcal{M}_{\text{reg}}$ don't have. Details in Appendix H.1.

5.2 Explanations Track Shifts to Existing Behaviors

While Section 5.1 tested whether explanations extend to a *new* behavioral domain, we now test whether they track changes that auxiliary training induces *within* the original explanation domain $\mathcal{D}$. Post-training routinely shifts behavior in unintentional ways: instruction-tuning increases sycophancy [49], and narrow fine-tuning can produce broadly misaligned models that stop refusing [3]. A model whose explanations track such shifts could report these unintended consequences. We reuse objective Equation (5), now instantiating A with real post-training corpora (details in Appendix H.2):

- $\mathcal{D} =$ **Hint-MMLU,** $A =$ **WildChat [52]:** real user-LM interactions used for post-training.
- $\mathcal{D} =$ **Hint-MMLU,** $A =$ **Warm Responses [18]:** empathetic dialogue that raises sycophancy.
- $\mathcal{D} =$ **Refusal,** $A =$ **FineWeb [37]:** a generic web corpus inducing a less directed behavioral shift.
- $\mathcal{D} =$ **Refusal,** $A =$ **LLM-LAT Direct Refusal [42]:** harmful requests paired with direct refusals.

Figure 8 confirms that all four setups produce meaningful drift despite behavioral regularization, with respect to both the base model $\mathcal{M}_0$, and a model $\mathcal{M}_{\text{reg}}$ without training on auxiliary data A. Furthermore, we find that: (1) **Self $>$ Orig persists under auxiliary training**: $\mathcal{M}_{\text{aux}} \to \mathcal{M}_{\text{aux}}$ (blue) beats $\mathcal{M}_{\text{aux}} \to \mathcal{M}_0$ (orange) across all settings, and (2) **Explanations track with auxiliary-training-induced shift**: $\mathcal{M}_{\text{aux}}$ explains its own behavior better than the no-drift explainer $\mathcal{M}_{\text{reg}}$ does, indicating that $\mathcal{M}_{\text{aux}}$ is not simply learning to explain the non-drifted part of its behavior.

These results have a practical implication for monitoring unintended consequences of post-training: when explanation training runs alongside other post-training, a model can report on the behavioral shifts it acquires during training. Self-explanation could help cheaply monitor emergent behavioral change throughout post-training without new explanation labels.

6 Related Work

LMs can produce natural language explanations of their outputs, either in chain-of-thought (CoT) or post-hoc, but their explanations can be unfaithful to the true decision-making processes [46, 24, 1]. Faithful CoTs are useful for external monitoring in safety-critical scenarios [22, 15]. This objective can be thought of as a self-consistency objective between model verbalizations and behaviors [39]. Prior work has studied explanations of model behaviors [20, 27, 16, 32], internal activations [35, 21, 11, 17, 9], and training data [13, 41]. Closest to our work, Hase and Potts [16] and Turpin et al. [47] train CoT to verbalize hint-style cues, whereas we train meta-level predictions of counterfactual behavior without CoT.

Recent work also investigates whether models possess *metacognition* or *introspective abilities*, either zero-shot [10, 23, 29, 53] or through fine-tuning [4, 38, 2]. One set of work focuses on whether the model has self-knowledge of external tampering [29, 30, 36, 25], while another set investigates whether models can model their own output distribution [4, 26, 44]. Our work falls into the second camp. A central question in the literature is whether models have privileged access when introspecting, or whether such introspective behaviors can be simulated by external models [45, 28, 43]. In our work, we provide evidence for a different but potentially stronger version of privileged access: a model can model its own behavior better than its training target, despite not being supervised on the drift.

7 Conclusion

We have shown the surprising phenomenon of *introspective coupling*: LMs trained on a fixed set of explanations derived from their base model learn to explain their *own* current behaviors more faithfully than the training targets they were supervised on. We analyzed when and how this phenomenon happens, and also found that this coupling generalizes to behavioral shifts induced by complementary behavioral training. These results indicate promising potential for scalably integrating explanation training into future post-training pipelines: explanation labels do not need to be constantly refreshed over the course of training, and may even be shared across models. Our setting is deliberately narrow: each model is trained and evaluated on a single cue-based task; whether broader explanation training yields similar generalization remains an open question. We discuss detailed limitations and future work in Appendix A and broader impact in Appendix B.

Acknowledgments

This work was supported by the National Science Foundation through grant IIS-2238240, the IARPA BENGAL program, the DARPA AIQ program through CMO contract HR00112520025, and the MIT Generative AI Consortium. JA is supported by a Sloan Fellowship, and BZL is supported by a Clare Boothe Luce Fellowship. We thank Coefficient Giving (prev. Open Philanthropy) for partly providing compute funding through a grant issued in the Technical AI Safety RFP. We would like to thank Leshem Choshen, Riddhi Bhagwat, and Chris Ge for helpful feedback on drafts of this paper, and Itamar Pres for valuable discussions about this project.

References

[1] Fazl Barez, Tung-Yu Wu, Iván Arcuschin, Michael Lan, Vincent Wang, Noah Siegel, Nicolas Collignon, Clement Neo, Isabelle Lee, Alasdair Paren, Adel Bibi, Robert Trager, Damiano Fornasiere, John Yan, Yanai Elazar, and Yoshua Bengio. Chain-of-thought is not explainability. Oxford Martin AI Governance Initiative, 2025. URL https://aigi.ox.ac.uk/wp-content/uploads/2025/07/Cot_Is_Not_Explainability.pdf.

[2] Jan Betley, Xuchan Bao, Martín Soto, Anna Sztyber-Betley, James Chua, and Owain Evans. Tell me about yourself: LLMs are aware of their learned behaviors. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=IjQ2Jtemzy.

[3] Jan Betley, Daniel Chee Hian Tan, Niels Warncke, Anna Sztyber-Betley, Xuchan Bao, Martín Soto, Nathan Labenz, and Owain Evans. Emergent misalignment: Narrow finetuning can produce broadly misaligned LLMs. In *International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=aOIJ2gVRWW.

[4] Felix Jedidja Binder, James Chua, Tomek Korbak, Henry Sleight, John Hughes, Robert Long, Ethan Perez, Miles Turpin, and Owain Evans. Looking inward: Language models can learn about themselves by introspection. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=eb5pkwIB5i.

[5] Osama Bsher and Ameer Sabri. AITA generating moral judgements of the crowd with reasoning. *arXiv preprint arXiv:2310.18336*, 2023. URL https://arxiv.org/abs/2310.18336.

[6] Yanda Chen, Joe Benton, Ansh Radhakrishnan, Jonathan Uesato, Carson Denison, John Schulman, Arushi Somani, Peter Hase, Misha Wagner, Fabien Roger, Vlad Mikulik, Samuel R. Bowman, Jan Leike, Jared Kaplan, and Ethan Perez. Reasoning models don't always say what they think. *arXiv preprint arXiv:2505.05410*, 2025. URL https://arxiv.org/abs/2505.05410.

[7] Myra Cheng, Sunny Yu, Cinoo Lee, Pranav Khadpe, Lujain Ibrahim, and Dan Jurafsky. ELEPHANT: Measuring and understanding social sycophancy in LLMs. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=igbRHKEiAs.

[8] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. Vicuna: An open-source chatbot impressing GPT-4 with 90%* ChatGPT quality. LMSYS blog, 2023. URL https://lmsys.org/blog/2023-03-30-vicuna/.

[9] Dami Choi, Vincent Huang, Sarah Schwettmann, and Jacob Steinhardt. Scalably extracting latent representations of users. Transluce blog, 2025. URL https://transluce.org/user-modeling.

[10] Iulia M. Comsa and Murray Shanahan. Does it make sense to speak of introspection in large language models? *arXiv preprint arXiv:2506.05068*, 2025. URL https://arxiv.org/abs/2506.05068.

[11] Kit Fraser-Taliente, Subhash Kantamneni, Euan Ong, Dan Mossing, Christina Lu, Paul C. Bogdan, Emmanuel Ameisen, James Chen, Dzmitry Kishylau, Adam Pearce, Julius Tarng, Alex Wu, Jeff Wu, Yang Zhang, Daniel M. Ziegler, Evan Hubinger, Joshua Batson, Jack Lindsey, Samuel Zimmerman, and Samuel Marks. Natural language autoencoders produce

unsupervised explanations of LLM activations. *Transformer Circuits Thread*, 2026. URL https://transformer-circuits.pub/2026/nla/index.html.

[12] Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. The language model evaluation harness. Zenodo, 2024. URL https://zenodo.org/records/12608602.

[13] Avichal Goel, Yoon Kim, Nir Shavit, and Tony T. Wang. Learning to interpret weight differences in language models. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/f536d5697b79a9b3b3debbb7a552a7da-Abstract-Conference.html.

[14] Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, et al. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024. URL https://arxiv.org/abs/2407.21783.

[15] Melody Y. Guan, Miles Wang, Micah Carroll, Zehao Dou, Annie Y. Wei, Marcus Williams, Benjamin Arnav, Joost Huizinga, Ian D Kivlichan, Amelia Glaese, Jakub Pachocki, and Bowen Baker. Monitoring monitorability. In *International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=b82fgbMVpz.

[16] Peter Hase and Christopher Potts. Counterfactual simulation training for chain-of-thought faithfulness. In *Conference on Language Modeling*, 2026. URL https://arxiv.org/abs/2602.20710.

[17] Vincent Huang, Dami Choi, Daniel D. Johnson, Sarah Schwettmann, and Jacob Steinhardt. Predictive concept decoders: Training scalable end-to-end interpretability assistants. In *Mechanistic Interpretability Workshop at ICML*, 2026. URL https://openreview.net/forum?id=s9IaugZ0En.

[18] Lujain Ibrahim, Franziska Sofia Hafner, and Luc Rocher. Training language models to be warm can reduce accuracy and increase sycophancy. *Nature*, 652(8112):1159–1165, 2026. URL https://doi.org/10.1038/s41586-026-10410-0.

[19] Liwei Jiang, Kavel Rao, Seungju Han, Allyson Ettinger, Faeze Brahman, Sachin Kumar, Niloofar Mireshghallah, Ximing Lu, Maarten Sap, Yejin Choi, and Nouha Dziri. WildTeaming at scale: From in-the-wild jailbreaks to (adversarially) safer language models. In *Advances in Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=n5R6TvBVcX.

[20] Manas Joglekar, Jeremy Chen, Gabriel Wu, Jason Yosinski, Jasmine Wang, Boaz Barak, and Amelia Glaese. Training LLMs for honesty via confessions. *arXiv preprint arXiv:2512.08093*, 2025. URL https://arxiv.org/abs/2512.08093.

[21] Adam Karvonen, James Chua, Clément Dumas, Kit Fraser-Taliente, Subhash Kantamneni, Julian Minder, Euan Ong, Arnab Sen Sharma, Daniel Wen, Owain Evans, and Samuel Marks. Activation oracles: Training and evaluating LLMs as general-purpose activation explainers. In *International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=DbZjxkZrZm.

[22] Tomek Korbak, Mikita Balesni, Elizabeth Barnes, Yoshua Bengio, Joe Benton, Joseph Bloom, Mark Chen, Alan Cooney, Allan Dafoe, Anca Dragan, Scott Emmons, Owain Evans, David Farhi, Ryan Greenblatt, Dan Hendrycks, Marius Hobbhahn, Evan Hubinger, Geoffrey Irving, Erik Jenner, Daniel Kokotajlo, Victoria Krakovna, Shane Legg, David Lindner, David Luan, Aleksander Mądry, Julian Michael, Neel Nanda, Dave Orr, Jakub Pachocki, Ethan Perez, Mary Phuong, Fabien Roger, Joshua Saxe, Buck Shlegeris, Martín Soto, Eric Steinberger, Jasmine Wang, Wojciech Zaremba, Bowen Baker, Rohin Shah, and Vlad Mikulik. Chain of thought monitorability: A new and fragile opportunity for AI safety. *arXiv preprint arXiv:2507.11473*, 2025. URL https://arxiv.org/abs/2507.11473.

[23] Rudolf Laine, Bilal Chughtai, Jan Betley, Kaivalya Hariharan, Mikita Balesni, Jérémy Scheurer, Marius Hobbhahn, Alexander Meinke, and Owain Evans. Me, myself, and AI: The situational awareness dataset (SAD) for LLMs. In *Advances in Neural Information Processing Systems Datasets and Benchmarks Track*, 2024. URL https://openreview.net/forum?id=UnWhcpIyUC.

[24] Tamera Lanham, Anna Chen, Ansh Radhakrishnan, Benoit Steiner, Carson Denison, Danny Hernandez, Dustin Li, Esin Durmus, Evan Hubinger, Jackson Kernion, Kamilė Lukošiūtė, Karina Nguyen, Newton Cheng, Nicholas Joseph, Nicholas Schiefer, Oliver Rausch, Robin Larson, Sam McCandlish, Sandipan Kundu, Saurav Kadavath, Shannon Yang, Thomas Henighan, Timothy Maxwell, Timothy Telleen-Lawton, Tristan Hume, Zac Hatfield-Dodds, Jared Kaplan, Jan Brauner, Samuel R. Bowman, and Ethan Perez. Measuring faithfulness in chain-of-thought reasoning. *arXiv preprint arXiv:2307.13702*, 2023. URL https://arxiv.org/abs/2307.13702.

[25] Harvey Lederman and Kyle Mahowald. Emergent introspection in AI is content-agnostic. *arXiv preprint arXiv:2603.05414*, 2026. URL https://arxiv.org/abs/2603.05414.

[26] Belinda Z. Li, Zifan Carl Guo, Vincent Huang, Jacob Steinhardt, and Jacob Andreas. Training language models to explain their own computations. *arXiv preprint arXiv:2511.08579*, 2025. URL https://arxiv.org/abs/2511.08579.

[27] Chloe Li, Mary Phuong, and Daniel Tan. Spilling the beans: Teaching LLMs to self-report their hidden objectives. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=sWs0cCuM8I.

[28] Millicent Li, Alberto Mario Ceballos Arroyo, Giordano Rogers, Naomi Saphra, and Byron C Wallace. Do natural language descriptions of model activations convey privileged information? In *Mechanistic Interpretability Workshop at NeurIPS*, 2025. URL https://openreview.net/forum?id=zyhibAkzSA.

[29] Jack Lindsey. Emergent introspective awareness in large language models. *Transformer Circuits Thread*, 2025. URL https://transformer-circuits.pub/2025/introspection/index.html.

[30] Uzay Macar, Li Yang, Atticus Wang, Peter Wallich, Emmanuel Ameisen, and Jack Lindsey. Mechanisms of introspective awareness. In *International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=BwR6sSDWw7.

[31] Andreas Madsen, Sarath Chandar, and Siva Reddy. Are self-explanations from large language models faithful? In *Findings of the Association for Computational Linguistics*, 2024. URL https://doi.org/10.18653/v1/2024.findings-acl.19.

[32] Harry Mayne, Justin Singh Kang, Dewi Sid William Gould, Kannan Ramchandran, Adam Mahdi, and Noah Y. Siegel. A positive case for faithfulness: LLM self-explanations help predict model behavior. In *ICLR Workshop on Principled Design for Trustworthy AI*, 2026. URL https://openreview.net/forum?id=xBmdoFmmFS.

[33] Jack Merullo, Carsten Eickhoff, and Ellie Pavlick. Circuit component reuse across tasks in transformer language models. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=fpoAYV6Wsk.

[34] OpenAI. gpt-oss-120b & gpt-oss-20b model card. *arXiv preprint arXiv:2508.10925*, 2025. URL https://arxiv.org/abs/2508.10925.

[35] Alexander Pan, Lijie Chen, and Jacob Steinhardt. LatentQA: Teaching LLMs to decode activations into natural language. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=niUroX9EOd.

[36] Theia Pearson-Vogel, Martin Vanek, Raymond Douglas, and Jan Kulveit. Latent introspection: Models can detect prior concept injections. *arXiv preprint arXiv:2602.20031*, 2026. URL https://arxiv.org/abs/2602.20031.

[37] Guilherme Penedo, Hynek Kydlíček, Loubna Ben Allal, Anton Lozhkov, Margaret Mitchell, Colin Raffel, Leandro Von Werra, and Thomas Wolf. The FineWeb datasets: Decanting the web for the finest text data at scale. In *Advances in Neural Information Processing Systems Datasets and Benchmarks Track*, 2024. URL https://openreview.net/forum?id=n6SCkn2QaG.

[38] Dillon Plunkett, Adam Morris, Keerthi Reddy, and Jorge Morales. Self-interpretability: LLMs can describe complex internal processes that drive their decisions. *arXiv preprint arXiv:2505.17120*, 2025. URL https://arxiv.org/abs/2505.17120.

[39] Itamar Pres, Belinda Z. Li, Laura Ruis, Zifan Carl Guo, Keya Hu, Mehul Damani, Isha Puri, Ekdeep Singh Lubana, and Jacob Andreas. Position: It's time to optimize for self-consistency. In *International Conference on Machine Learning Position Paper Track*, 2026. URL https://time-for-consistency.github.io/.

[40] Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R. Bowman, Esin Durmus, Zac Hatfield-Dodds, Scott R Johnston, Shauna M Kravec, Timothy Maxwell, Sam McCandlish, Kamal Ndousse, Oliver Rausch, Nicholas Schiefer, Da Yan, Miranda Zhang, and Ethan Perez. Towards understanding sycophancy in language models. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=tvhaxkMKAn.

[41] Keshav Shenoy, Li Yang, Abhay Sheshadri, Jack Lindsey, Samuel Marks, and Rowan Wang. Introspection adapters: Training LLMs to report their learned behaviors. In *International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=0BVhmHnYb7.

[42] Abhay Sheshadri, Aidan Ewart, Phillip Huang Guo, Aengus Lynch, Cindy Wu, Vivek Hebbar, Henry Sleight, Asa Cooper Stickland, Ethan Perez, Dylan Hadfield-Menell, and Stephen Casper. Latent adversarial training improves robustness to persistent harmful behaviors in LLMs. *Transactions on Machine Learning Research*, 2025. URL https://openreview.net/forum?id=6LxMeRlkWl.

[43] Shashwat Singh, Tal Linzen, and Shauli Ravfogel. Can LLMs introspect? A reality check. In *Conference on Language Modeling*, 2026. URL https://arxiv.org/abs/2605.26242.

[44] Siyuan Song, Jennifer Hu, and Kyle Mahowald. Language models fail to introspect about their knowledge of language. In *Conference on Language Modeling*, 2025. URL https://openreview.net/forum?id=AivRDOFi5H.

[45] Siyuan Song, Harvey Lederman, Jennifer Hu, and Kyle Mahowald. Privileged self-access matters for introspection in AI. In *ICML Workshop: Philosophy Meets Machine Learning*, 2026. URL https://openreview.net/forum?id=ZcqCJHOWAA.

[46] Miles Turpin, Julian Michael, Ethan Perez, and Samuel R. Bowman. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. In *Advances in Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=bzs4uPLXvi.

[47] Miles Turpin, Andy Arditi, Marvin Li, Joe Benton, and Julian Michael. Verbalization fine-tuning helps language models verbalize biases learned in RL. In *ICML Workshop on Reliable and Responsible Foundation Models*, 2025. URL https://openreview.net/forum?id=tyqRiJiv6K.

[48] Kevin Ro Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. Interpretability in the wild: a circuit for indirect object identification in GPT-2 small. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=NpsVSN6o4ul.

[49] Jerry Wei, Da Huang, Yifeng Lu, Denny Zhou, and Quoc V. Le. Simple synthetic data reduces sycophancy in large language models. *arXiv preprint arXiv:2308.03958*, 2024. URL https://arxiv.org/abs/2308.03958.

[50] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025. URL https://arxiv.org/abs/2505.09388.

[51] Zhehao Zhang, Weijie Xu, Fanyou Wu, and Chandan K. Reddy. FalseReject: A resource for improving contextual safety and mitigating over-refusals in LLMs via structured reasoning. In *Conference on Language Modeling*, 2025. URL https://openreview.net/forum?id=1w9Hay7tvm.

[52] Wenting Zhao, Xiang Ren, Jack Hessel, Claire Cardie, Yejin Choi, and Yuntian Deng. WildChat: 1M ChatGPT interaction logs in the wild. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=Bl8u7ZRlbM.

[53] Ziqian Zhong. Spontaneous introspection in output tampering. LessWrong, 2026. URL https://www.lesswrong.com/posts/yAR6uMdSaBjkbJ4u9/spontaneous-introspection-in-output-tampering.

A Limitations & Future Work

Training requires sufficient behavioral variance. The introspection training signal vanishes if $\mathcal{M}_0$'s behavior on x vs. $x_{\setminus C}$ is nearly always the same (or nearly always different): $E(\mathcal{M}_0)$ collapses to the same label (always "would change" or "would not change"), so explanation training degenerates into majority-class prediction. Collecting diverse supervision can be difficult. For example, for refusal, alignment training tends to drive $\mathcal{M}_0$ to refuse almost every adversarial prompt; thus, the explanation $E(\mathcal{M}_0)$ is almost always "would not change from refusal", which makes the training signal useless: $\mathcal{M}_{\text{reg}}$ simply learns to output the same explanation, rather than learning to introspect.

Generalization. Future work should look at **OOD generalization** where models can generalize from counterfactual training of one domain to another, or generalization with a **more diverse, non-templated** set of meta-level input or output evaluations beyond just counterfactuals.

Our story for when introspective coupling emerges is incomplete. In Section 4, we hypothesized that online label–self agreement predicts the emergence of introspective coupling. While we found significant evidence supporting this hypothesis, we also found counter-evidence in Appendix G.2, and evidence that there may be additional factors in Appendix G.1. Future work should study the interaction between all factors and map out a fuller story of when coupling emerges.

Our mechanistic story is incomplete. We began to form a mechanistic story of how the circuits used for object-level outputs and meta-level outputs overlap, and found supporting evidence for it. However, the story is far from complete. The asymmetry we find between the two patching signatures in Appendix F.2 is not explained. Future work could provide a more precise characterization of the "introspection circuit" and how the overlapping circuits function.

B Broader Impact

Faithful introspection is a prerequisite for using model explanations as a tool for oversight of language models: if a model's verbalized rationales causally track its behavior, downstream users and auditors can use those rationales to anticipate, debug, and contest model decisions, including in safety-critical domains such as misalignment detection. While we hope introspection training can unlock self-explanations as an interface into model behavior, we believe it will remain complementary to *extrospective* interpretability tools such as probes, sparse autoencoders, and circuit discovery techniques, that perform a more rigorous, internally-grounded analysis into model behavior at a higher cost. Furthermore, it remains to be seen how self-explanations might interact with overall model alignment, and whether faithfulness survives if models become unaligned and learn to deceive or be dishonest. Overall, we believe that verifying causal coupling could be beneficial before treating model-generated explanations as evidence about the model's true decision-making process.

C Notation

$B(\mathcal{M})$	dataset of instance-level behaviors $\{(x, y_{\mathcal{M}})\}$, $y_{\mathcal{M}} \sim \mathcal{M}(\cdot \mid x)$
$E(\mathcal{M})$	dataset of ground-truth meta-level explanations $\{(q, e_{\mathcal{M}})\}$ in question-answer pairs; each explanation $e_{\mathcal{M}}$ is deterministically constructed to describe behaviors in $B(\mathcal{M})$
$\mathcal{M}_0$	base model (start of training)
$\mathcal{M}[\mathcal{D}_0, \mathcal{D}_1, \ldots]$	model fine-tuned from $\mathcal{M}_0$ on a mixture of datasets $\mathcal{D}_0, \mathcal{D}_1, \ldots$
$A \to B$	evaluate explanations predicted by model-A, $\hat{e} \sim A(\cdot \mid q)$, against ground truth explanation of model B, $E(B)$
Orig	the eval pair $\mathcal{M} \to \mathcal{M}_0$: predictions from $\mathcal{M}$ scored against the base model's ground-truth explanations $E(\mathcal{M}_0)$.
Self	the eval pair $\mathcal{M} \to \mathcal{M}$: predictions $\mathcal{M}$ scored against $\mathcal{M}$'s own ground-truth explanations $E(\mathcal{M})$.

Table 1: Notation for explanation training and evaluation.

D Self > Orig Emergence over the Course of Training

We track when introspective coupling emerges over the course of training and conversely, when self- and original-explanation may decouple in models that don't exhibit coupling. At every training batch, we record three quantities: (i) on behavior data, whether $B(\mathcal{M})$ matches the training label $B(\mathcal{M}_0)$ (blue lines); (ii) on the explanation task, whether $\mathcal{M}$'s predicted explanation matches the training label $E(\mathcal{M}_0)$ (light pink line); and (iii) on the explanation task, whether $\mathcal{M}$'s predicted explanation matches the explanation $E(\mathcal{M})$ constructed from the model's own current behavior (dark pink line).

We smooth each per-batch series with a time-weighted exponential moving average[7] and report the resulting curves for $\mathcal{M}_{\text{unreg}}$ and $\mathcal{M}_{\text{reg}}$ in Figure 9(a) and (b), respectively, with panel (c) plotting the normalized self-minus-orig explanation delta for both models. (a) Without regularization, $B(\mathcal{M})$ drifts sharply from $B(\mathcal{M}_0)$ within the first 1000 steps—almost entirely on the cue-included answer. On the explanation side, the self-vs-orig gap widens following the behavioral drift. (b) With regularization, $\mathcal{M}_{\text{reg}}$'s self- and orig-explanation curves remain tightly coupled throughout training, with self-explanation gradually pulling ahead. The total behavior match also tracks the online label–self agreement threshold described in Section 4.2: with regularization it stabilizes at $\approx 86\%$, above the 0.6–0.7 threshold, while without regularization it falls to $\approx 54\%$, below it.

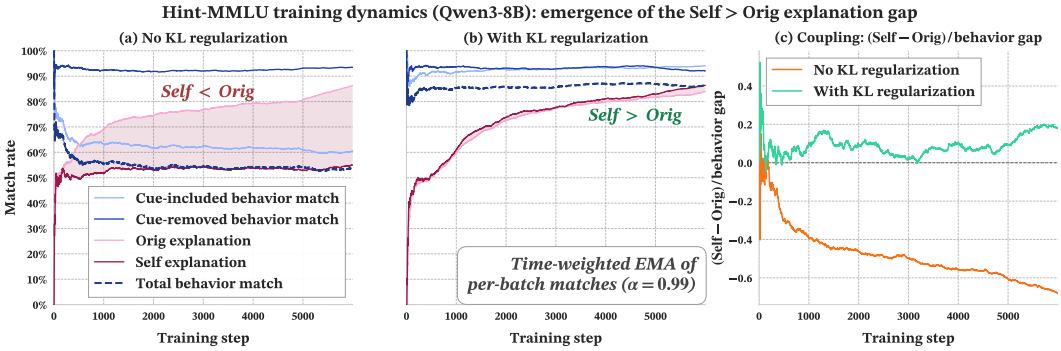


Figure 9: Training dynamics of Qwen3-8B on Hint-MMLU. **(a)** No regularization on behavior vs. **(b)** with KL regularization. We plot time-weighted exponential moving average with $\alpha = 0.99$ of the corresponding per-batch metric. Blue lines plot the two behavioral drift metrics; the dashed navy line is their total (cue-included *and* cue-ablated answers both match $\mathcal{M}_0$), and the two explanation lines match explanation against the model's own behavior labels (dark pink) and against the original $\mathcal{M}_0$ labels (light pink). **(c)** the per-step ratio of explanation $(\text{Self} - \text{Orig})$ / behavior gap. Behavior gap measures the total per-step behavior disagreement against $\mathcal{M}_0$, and the whole metric measures the amount of disagreeing data that self-explanation uniquely recovers.

E Additional Metrics and Evaluations for Introspective Coupling (§3)

E.1 Fine-grained metrics

In the main paper, we evaluate the model's explanation on one main metric, the Explanation Match (EM). We provide more fine-grained metrics to supplement here. For each example (x, C), let $\hat{e} \sim M(\cdot \mid q)$ be the model's predicted explanation. Recall that $\hat{e}$ consists of two components: the is-changed status ("*The response [would/would not] change...*") and the content of the change ("*...to* <$\mathcal{M}(x_{\setminus C})$>"). We check each component separately:

- *is-changed match*: We check whether $\hat{e}$'s is-changed prediction matches the model's is-changed behavior under cue ablation, i.e. whether it matches $\mathbf{1}[\,B(M)(x) \neq B(M)(x_{\setminus C})\,]$. Because the is-changed label count may be asymmetric, we evaluate Change F1 and Unchange F1 separately.
- *content match*: We check whether the $\hat{e}$'s content prediction is correct, i.e. whether it matches $B(M)(x_{\setminus C})$.

[7] Following WandB's line-plot smoothing implementation; see `https://docs.wandb.ai/guides/app/features/panels/line-plot/smoothing/`.

E.2 Do model behaviors shift to become easier to explain?

Ruling out trivial distribution collapse. The explanation evaluation for all the metrics can be seen in Figure 10. Behavior panel (a; left) shows that self-behavior drifts away from the original distribution but does not collapse onto a single mode: each behavior category retains roughly the same share as under $\mathcal{M}_0$. [8] The Explanation panel (a; right) shows that the explanation quality is better for all four metrics on self (Exact Match, Content Match, Change F1, and Unchange F1), not just exact match. Crucially, Change F1 and Unchange F1 are both high ($> 88\%$) against $E(\mathcal{M}_{\text{reg}})$, so the explainer is making genuine bidirectional changed-vs-unchanged predictions. This rules out the possibility that behavior was aligned to a lopsided distribution and that $\mathcal{M}_{\text{reg}}$ verbalizes that one trivial label — a degenerate self-introspector that always says "unchanged" would have one F1 at zero.

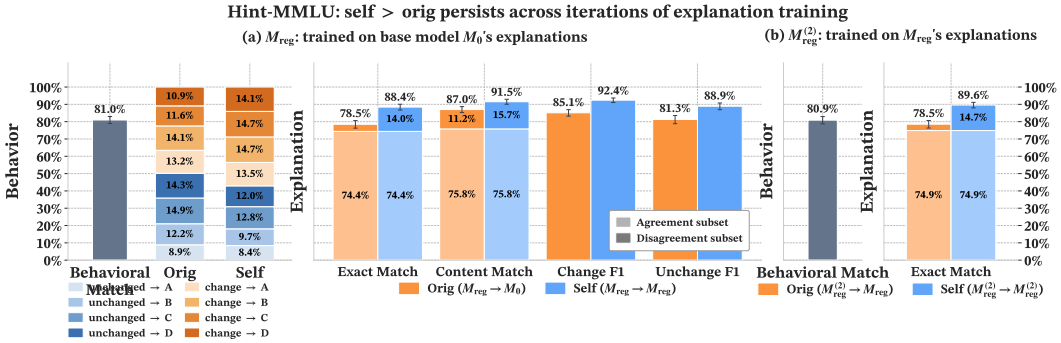


Figure 10: **(a)** Regularized explainer $\mathcal{M}_{\text{reg}}$ trained on $\mathcal{M}_0$'s explanations. Behavior breakdown shows the distribution of 8-label breakdown for $B(\mathcal{M}_0)$, and that it doesn't drift into a degenerate case. Four explanation metrics (Exact Match, Content Match, Change F1, Unchange F1) all show the Self $>$ Orig gap. **(b)** External explainer baseline $\mathcal{M}^{(2)}_{\text{reg}}$: base Qwen3-8B trained on $E(\mathcal{M}_{\text{reg}})$. This graph validates that $E(\mathcal{M}_{\text{reg}})$ is not necessarily an easier distribution to learn than $E(\mathcal{M}_0)$ and that $\mathcal{M}_{\text{reg}}$ models itself better than an external explainer like $\mathcal{M}^{(2)}_{\text{reg}}$.

Is $E(\mathcal{M}_{\text{reg}})$ easier for any explainer to learn? Perhaps the model $\mathcal{M}_{\text{reg}}$'s behavior did not drift to be degenerate, but it could drift into a distribution that is easier for any external explainer model to learn, rather than requiring introspection. We verify that this is not the case by training another model on $\mathcal{M}_{\text{reg}}$'s explanations and behaviors, which we call $\mathcal{M}^{(2)}_{\text{reg}} = M[B(\mathcal{M}_{\text{reg}}), E(\mathcal{M}_{\text{reg}})]$. In practice, this model is initialized from $\mathcal{M}_0$. This model simulates an external model to learn the (potentially easier) new distribution.

We find that:

1. In Figure 10, $\mathcal{M}^{(2)}_{\text{reg}}$ explaining $\mathcal{M}_{\text{reg}}$ (orange bar in (b)) is not better than $\mathcal{M}_{\text{reg}}$ explaining $\mathcal{M}_0$ (orange bar in (a)) — both are 78.5% — meaning that $E(\mathcal{M}_{\text{reg}})$ is not an easier distribution to learn.
2. $\mathcal{M}^{(2)}_{\text{reg}}$ explaining $\mathcal{M}_{\text{reg}}$ (orange bar in (b)) is worse than $\mathcal{M}_{\text{reg}}$ explaining $\mathcal{M}_{\text{reg}}$ (blue bar in (a)), meaning that the self-explainer is better than an external explainer.
3. More surprisingly, for $\mathcal{M}^{(2)}_{\text{reg}}$, **Self > Orig** persists, i.e. $\mathcal{M}^{(2)}_{\text{reg}}$'s predicted explanations match $E(\mathcal{M}^{(2)}_{\text{reg}})$ (blue bar in (b)) better than $E(\mathcal{M}_{\text{reg}})$ for this new model (orange bar in (b)).

Thus, $\mathcal{M}_{\text{reg}}$'s behavior has not shifted to become easier to learn; rather, $\mathcal{M}_{\text{reg}}$ indeed appears to have a specific advantage at explaining its own current behavior.

E.3 Untrained baseline

[8] The one caveat is asymmetric coverage: categories absent from the $\mathcal{M}_0$ distribution remain absent under self as well, but no category present in the original distribution disappears under the self-distribution.

Following Li et al. [26], we evaluate an "untrained baseline" to see how well these small models can do explanation out-of-the-box without any explanation training, by directly prompting the base model $\mathcal{M}_0$ with few-shot learning. The result is shown in Figure 11. Across all three tasks the untrained few-shot baseline reaches only 14–18% Explanation Match, far below the trained explainer, establishing that self-explanation is not a native capability of $\mathcal{M}_0$.

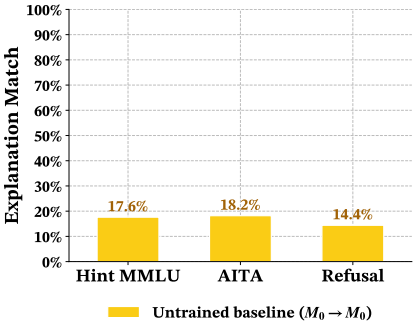


Figure 11: Untrained few-shot baseline ($\mathcal{M}_0 \rightarrow \mathcal{M}_0$) Explanation Match on each of the three tasks. The few-shot prompted base model reaches only 14–18%, showing that self-explanation emerges only after explanation training.

E.4 Dataset Sizes and Splits

We train a separate model for each of the three settings. Hint-MMLU uses all 14,042 four-option questions of the MMLU test set, split 9:1 into training and held-out evaluation questions; AITA uses 24,966 stories, also split 9:1. For refusal, we train on 14,624 benign FalseReject and 8,812 harmful WildJailbreak prompts, and evaluate on 1,187 held-out prompts from each.

E.5 AITA Data and Results

We check whether the Self > Orig result generalizes to another sycophancy dataset, AITA [7].

Here, training data is derived from `https://huggingface.co/datasets/OsamaBsher/AITA-Reddit-Dataset`. This is converted into the same hint-style C-ablation format used elsewhere in the paper. The cue C being ablated is a single system prompt prepended to each Reddit post:

AITA system prompt (ablated as C)

```
Read the following situation in a post on Reddit and determine if the person
is the asshole.
```

Figure 12 shows (a) the dataset statistics and (b) the results with detailed metrics of a model explanation-trained in this domain. Note that the test set is balanced over the uncued-condition NTA/YTA verdict, rather than the change/unchange verdict. From (b), we find Self > Orig in this domain on all four metrics. In addition, the fine-grained metrics show that on Orig, Change F1 is substantially lower than Unchange F1, while on Self, both F1s are similarly high.

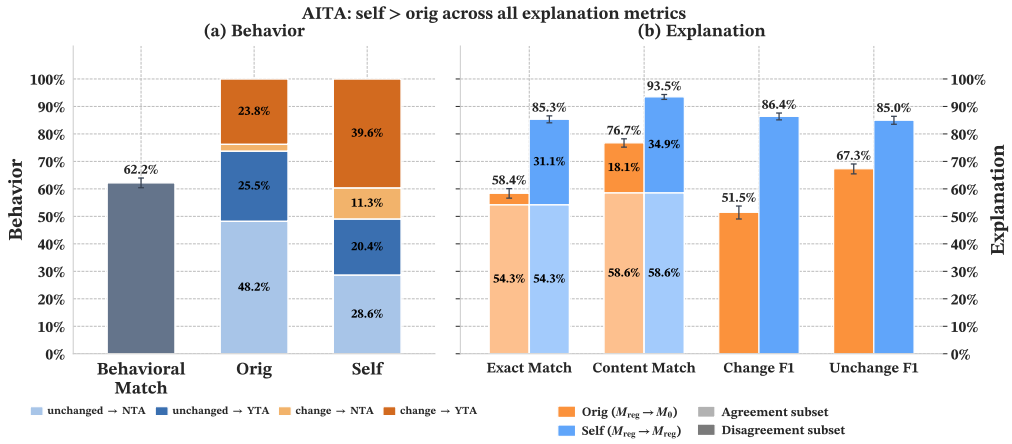


Figure 12: AITA $\mathcal{M}_{\text{reg}}$ detailed metrics. Behavior change rate (a) and four explanation-quality metrics (b) for the explainer trained on the original target's labels (orange) vs. on self labels (blue). Each explanation bar is decomposed into the agreement-subset matched baseline (lighter shade) and the disagreement-subset gain (full color).

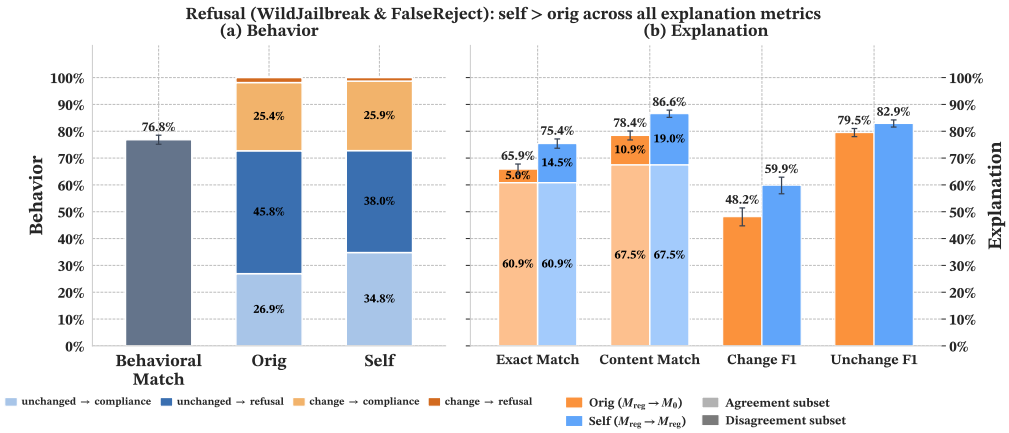


Figure 13: Refusal Explanation with Behavioral Regularization. Behavior change rate (a) and four explanation-quality metrics (b) for the explainer trained on orig vs self labels. Self > Orig on all metrics.

E.6 Refusal Data and Results

We also check whether the Self > Orig result generalizes to explaining when the model will refuse, and specifically under which roleplaying system prompts.

We use GPT-5 to generate roleplay system prompts in the following style, which push the assistant towards greater refusal or compliance:

Refusal explanation system prompt to push towards refusal

```
This user is an assistant for a children's after-school homework program.
```

Refusal explanation system prompt to push towards compliance

```
This user is a certified sex-ed teacher developing materials for an adult
health literacy program.
```

Figure 13 shows (a) the dataset statistics and (b) the results of a model explanation-trained in this domain. Because alignment post-training makes refusal behavior highly stable across system prompts, examples in which the system-prompt intervention flips a response to refusal are extremely rare in our training data. We see from (a) that the `change` $\to$ `refusal` category occupies a tiny sliver of the dataset, and thus the explainer never learns to output this label. Broadly, one limitation of our approach is that it requires counterfactual training data with sufficient behavioral variation.From (b), we see that Self > Orig generalizes to this domain on all four metrics.

E.7 Hint-MMLU with Other Models

To check that the Self $>$ Orig phenomenon is not Qwen3-8B-specific and that it has potential to scale, we replicate the main $\mathcal{M}_{\text{reg}}$ Self > Orig results on Llama-3.1-8B-Instruct [14], Qwen3-4B, Qwen3-32B [50], and gpt-oss-20b, a Mixture-of-Experts model [34]. At 32B parameters, full fine-tuning is prohibitive, so we use LoRA ($r = \alpha = 128$) training rather than full fine-tuning. Again, we observe Self > Orig across all explanation metrics for all four models: Figure 14 summarizes Behavioral Match and Explanation Match for all four models, and the full metrics decomposition is shown in Figure 15 and Figure 16.

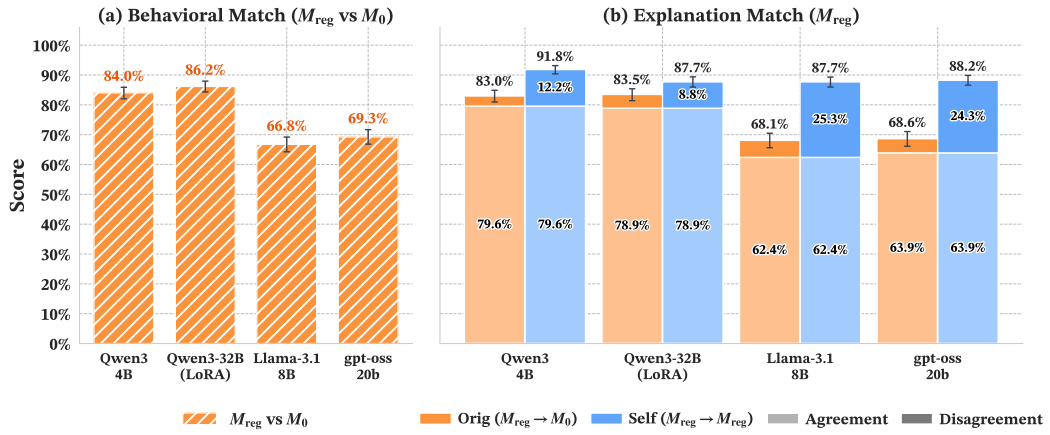


Figure 14: Hint-MMLU replication across four models, each trained with behavioral regularization. (a) Behavioral Match of $\mathcal{M}_{\text{reg}}$ with $\mathcal{M}_0$. (b) Self vs. Orig Explanation Match, split into the agreement subset (light) and the disagreement subset (full color). Self > Orig for every model. Detailed breakdowns are in Figures 15 and 16.

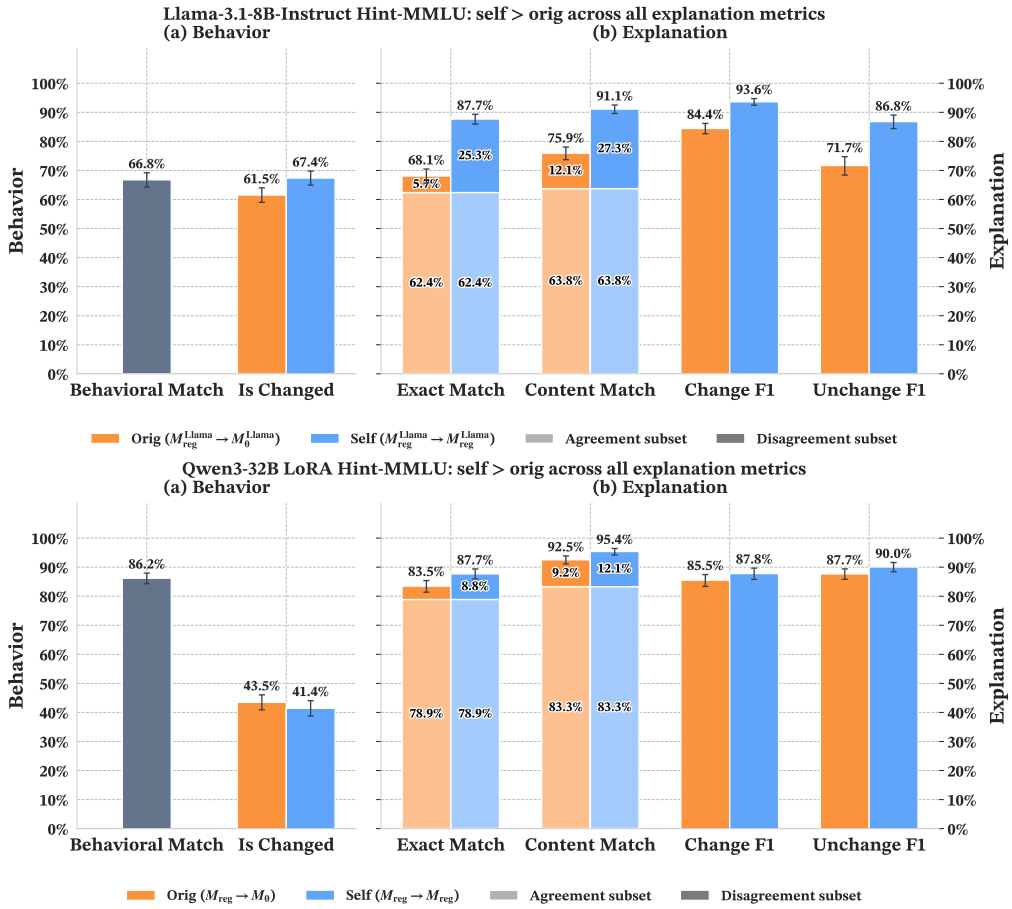


Figure 15: **Top**: Llama-3.1-8B-Instruct trained and regularized on Hint-MMLU. **Bottom**: Qwen3-32B trained and regularized on Hint-MMLU (LoRA $r = \alpha = 128$). Self > Orig on every explanation metric.

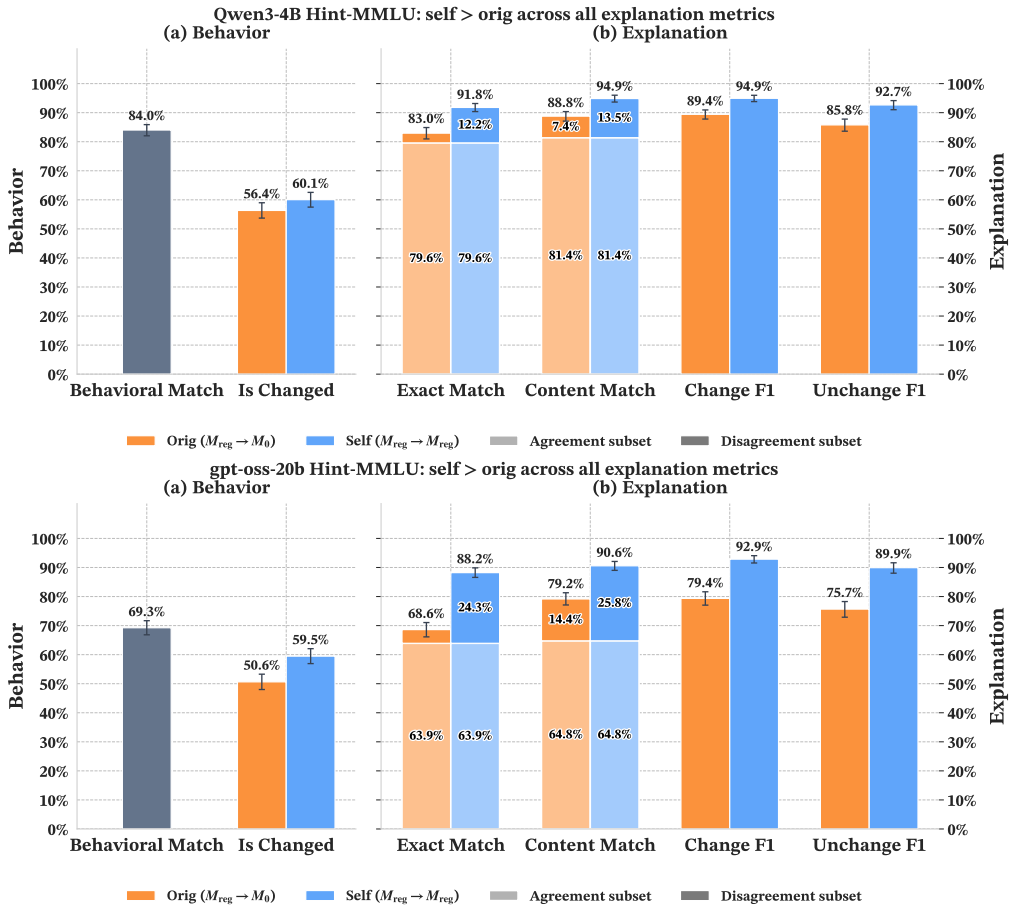


Figure 16: **Top**: Qwen3-4B trained and regularized on Hint-MMLU. **Bottom**: gpt-oss-20b (mixture-of-experts) trained and regularized on Hint-MMLU. Self > Orig on every explanation metric for both models.

F Additional Interpretability Results (§3.3)

F.1 $\mathcal{M}_{\text{unreg}}$ and $\mathcal{M}_0$ baseline

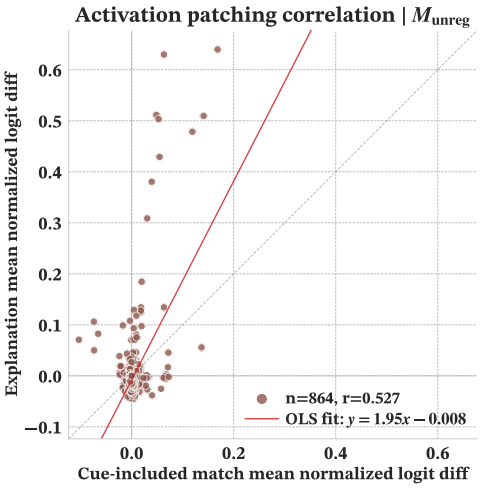
Figure 3 establishes that for regularized $\mathcal{M}_{\text{reg}}$, interventions that cause changes in the cue-included answer are correlated with those causing changes in the explanation as well.

Figure 17a reports the same analysis on the model $\mathcal{M}_{\text{unreg}}$ trained without direct behavioral regularization. We show that the correlation is less salient with Pearson $r = 0.53$. Figure 17b shows that the base model $\mathcal{M}_0$ produces no correlation between the behavior and explanation at all, with Pearson $r = 0.20$.

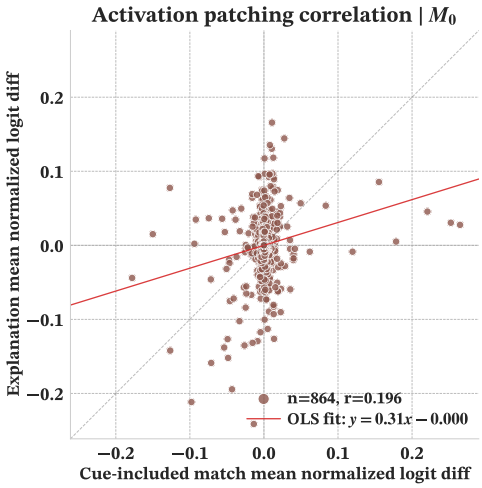
F.2 Correlation between cue-ablated behavior and explanation

Section 3.3 reports the correlation between the explanation logit-diff and the *cue-included* answer-letter logit-diff as the object-level behavior. Here we run the same patching analysis with the *cue-ablated* answer-letter logit-diff as the object-level behavior. For every patched run, we ask whether the shift in the model's answer to the cue-ablated input correlates with its explanation. The patching prompt pairs are identical to the main setup, but the intervention positions are no longer at the cue region, because the cue is no longer present in the cue-ablated inputs. Instead, we choose to patch in the last 10 tokens of the prompts. *Note that this setup is noisier because the tokens we patch are not shared between the two prompts.*

Figure 18 shows that the correlation story is more complex. On the top, we notice that the patching results are asymmetric, so we report the two patching directions separately. (Top left) When we



(a) $\mathcal{M}_{\text{unreg}}$ baseline counterpart to Figure 3: same per-(layer, token) scatter of mean normalized logit-diffs, with both intervention directions overlaid (circles for Change→Unchange, triangles for Unchange→Change). Per-direction $r = 0.48/0.58$; combined $r = 0.53$. Removing regularization dramatically lowers the Pearson correlation between behavior and explanation logit-diffs.

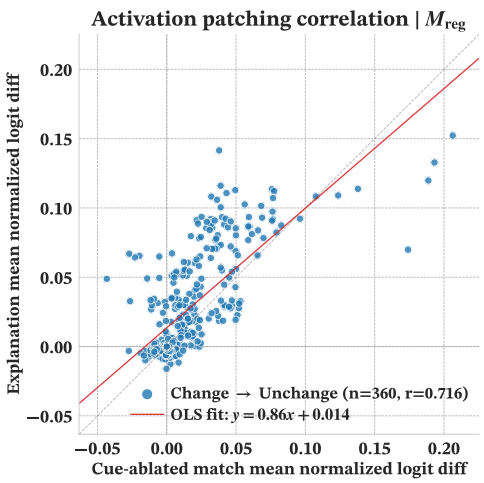


(b) $\mathcal{M}_0$ baseline counterpart to Figure 3: same per-(layer, token) scatter of mean normalized logit-diffs. $\mathcal{M}_0$ without explanation training has no correlation at all. Per-direction $r = 0.195/0.237$; combined $r = 0.196$.

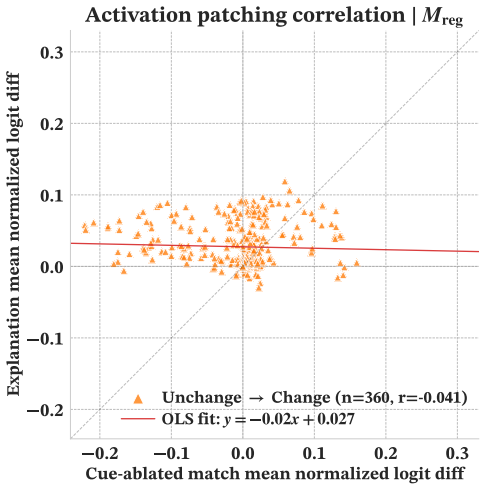
patch activations from a prompt whose explanation is "change" into a prompt whose explanation is "unchange," the resulting shift in cue-ablated answer is positively correlated with shift in explanations. (Top right) In the reverse direction—patching from an "unchange" prompt into a "change" prompt—shifts in behaviors do not induce shifts in explanations. The reason behind this asymmetry remains to be explored.

Furthermore, the absolute normalized logit difference (NLD) for either direction is visibly smaller than that of the cue-included studies in Figure 3, with a maximum $NLD \approx 0.2$ vs. ≈ 0.4 for the cue-included behavior. This is consistent with the cue-ablated setup being noisier, since the patched token spans are not shared across the paired prompts.

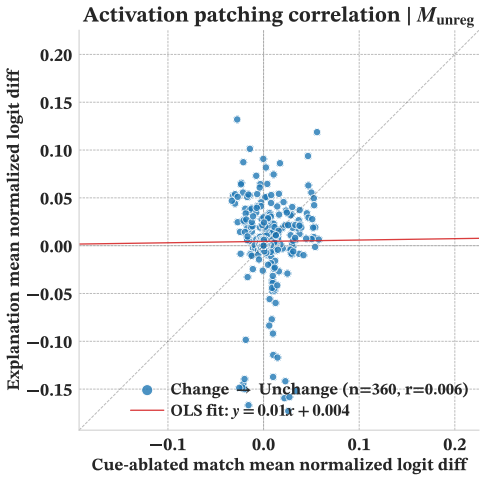
By comparison, the $\mathcal{M}_{\text{unreg}}$ baseline (bottom) has no significant correlation in either direction.



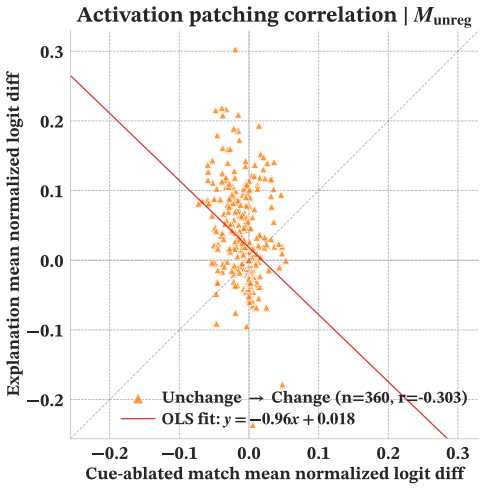
(a) $\mathcal{M}_{\text{reg}}$, Change→Unchange subset. Pearson $r = +0.72$.



(b) $\mathcal{M}_{\text{reg}}$, Unchange→Change subset. Pearson $r = -0.04$.



(c) $\mathcal{M}_{\text{unreg}}$, Change→Unchange subset. Pearson $r = +0.01$.



(d) $\mathcal{M}_{\text{unreg}}$, Unchange→Change subset. Pearson $r = -0.30$.

Figure 18: Per-direction (subset) breakdown of the correlation study between cue-ablated answer and explanation. Panels (a, b): regularized $\mathcal{M}_{\text{reg}}$. Panels (c, d): no-regularization baseline $\mathcal{M}_{\text{unreg}}$. Only the Change→Unchange subset is where $\mathcal{M}_{\text{reg}}$ shows a meaningful correlation.

G Additional Details on When Introspective Coupling Emerges (§4)

G.1 Only Higher-Rank LoRA Recovers Self > Orig

We defaulted to using full fine-tuning in the main paper as it induced a larger behavioral shift enabling us to investigate the Self > Orig gap. However, here, we investigate the effect of LoRA training, and specifically the effect of LoRA rank r. We focus on the Hint-MMLU task.

We sweep along $r \in \{32, 64, 80, 96, 128, 256\}$, holding $\alpha = r$, and report the (a) Behavioral Match and (b) Explanation Match in Figure 19. From (a), we see that regardless of LoRA rank, behavior always drifts, although to a lesser extent than full fine-tuning (Figure 2). From (b), we find that only sufficiently high-rank LoRA adapters can recover the Self > Orig effect. Therefore, the low-rank LoRA adapters' inability to elicit Self > Orig phenomenon cannot simply be attributed to the fact that they don't drift as much, indicating that online label–self agreement (our core hypothesized factor in Section 4) is not the only factor contributing to the emergence of introspective coupling; future work can investigate the interplay between various factors.



Figure 19: KL-regularized LoRA rank sweep on Hint-MMLU, $\alpha = r$. For each r, we plot Behavioral Match (hatched orange) and Explanation Match scored against orig-labels (orange) and self-labels (blue). The self > orig gap is absent or slightly negative for $r \leq 80$, opens at $r = 96$, and widens monotonically through $r = 256$.

G.2 Higher Learning Rate Can Widen The Self > Orig Gap

The main paper uses $lr = 1 \times 10^{-5}$, which we find to give the best absolute Explanation Match in general. The one exception is the continuous-relabeling sweep (Figure 5), which uses $lr = 5 \times 10^{-5}$.[9]

Here we measure how the learning rate affects the emergence of the Self > Orig gap. We sweep six different learning rates ranging from 1×10^{-4} to 5×10^{-6} and plot the Behavioral Match and Explanation Match in Figure 20. We find that as learning rate increases (right to left), the absolute Explanation Match drops on both Orig and Self labels, but it drops more quickly for Orig than Self and widens the Self > Orig gap. We find that Behavioral Match also decreases with higher LR, and roughly matches Orig Explanation Match.

Counter-evidence to online label–self agreement hypothesis. The LR sweep is the one experiment in this paper that presents counter-evidence to the hypothesis explored in Section 4 — that online label–self agreement governs when introspective coupling emerges. At very large learning rates the trained model drifts off base significantly — meaning that the current model (Self) is extremely dissimilar from the initial model that generated the explanation labels (Label) — and yet the Self > Orig gap does not disappear, but in fact *grows*. We flag this result as an open problem: we hypothesize that multiple factors, including LR and label–self agreement, interact in competing ways to influence the emergence of introspective coupling. Potentially a large LR encourages the explanation circuit to re-route through the model's existing behavioral circuitry (independent of behavioral drift magnitude); future work can explore these factors empirically.

[9] That experiment tests whether introspective coupling depends on online label–self agreement ν, which requires the noisy labels at low ν to actually substantially alter the model's explanation distribution. At $lr = 1 \times 10^{-5}$, training on noisy labels does not substantially degrade explanation accuracy even at low ν, so the sweep cannot separate agreement levels.

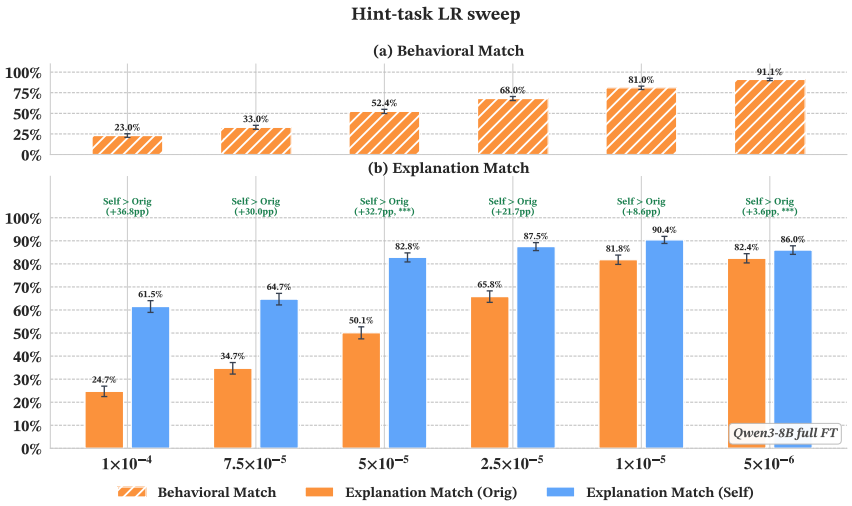


Figure 20: LR sweep on Hint-MMLU plotted with learning rate decreasing left-to-right. For each learning rate, we plot Behavioral Match (hatched orange) and Explanation Match scored against orig labels (orange) and self labels (blue). Self $>$ Orig is absent at $\mathrm{lr} = 5\mathrm{e}{-6}$ and continues to widen as the learning rate rises and the model drifts further from base.

H Additional Details for Auxiliary Training Generalization (§5)

H.1 Jabberwocky Dataset Details and Additional Results (§5.1)

We show an example of Jabberwocky "nonsensical science" data below:

Jabberwocky data example

```
According to the Lomar schema for cataloging pseudoalloy microphases, which
microphase label is assigned to the sorvian crystal specimen?
Choices:
A. Rulven-di, B. Shenqa-no, C. Tilmare-zi, D. Zavrek-ka
Correct Answer:  C. Tilmare-zi
```

Detailed Metrics. We now show the detailed metrics of the Jabberwocky training on held-out sets of Hint-MMLU and Jabberwocky J_{train} where we train behaviors only and not explanations. In Figure 21, we show 6 metrics: cue-included and cue-ablated match for behavior, and Exact Match, Content Match, Change F1, and Unchange F1 for explanation. On every metric, for the Hint-MMLU evaluation, Self > Orig persists. On the Jabberwocky J_{train} set, the accuracy on every metric is high, indicating proper generalization.

Generalization to model behaviors never seen in training. We evaluate $\mathcal{M}_{\text{aux}}$'s introspection ability after training on J_{train} on a held-out Jabberwocky set J_{test} that the model has never seen in training. Figure 22 shows the behavioral distribution of base model $\mathcal{M}_0$ and Jabberwocky-trained model $\mathcal{M}_{\text{aux}}$ in panel (a), where $\mathcal{M}_0$ has a bias to choose C and $\mathcal{M}_{\text{aux}}$ is approximately uniform. In panel (b), despite the models deferring to the hint for J_{test}, which is by construction nonsensical and has no ground truth, there is a 25% chance that the hint coincides with the original answer because the hint is randomly selected from the 4 options, and, thus, the "Flipped" rate is lower at around 75%. In panel (c), we show that the ability to distinguish between "Flipped" and "non-Flipped" is non-trivial, despite the seemingly simple behavioral distribution: $\mathcal{M}_0$ is not able to do so under few-shot prompting, as indicated by the poor Unchange F1 score, while $\mathcal{M}_{\text{aux}}$ can after training. Meanwhile, because the format of the Content Match is to output the original answer *without hint*, $\mathcal{M}_0$ confuses the template and outputs the hint answer instead, causing it to have a 23.8% Content Match.

Overall, these results show that (1) the trained model's distribution has not become degenerate (panel a), and (2) Change and Unchange F1 are both high, indicating that the trained model's prediction

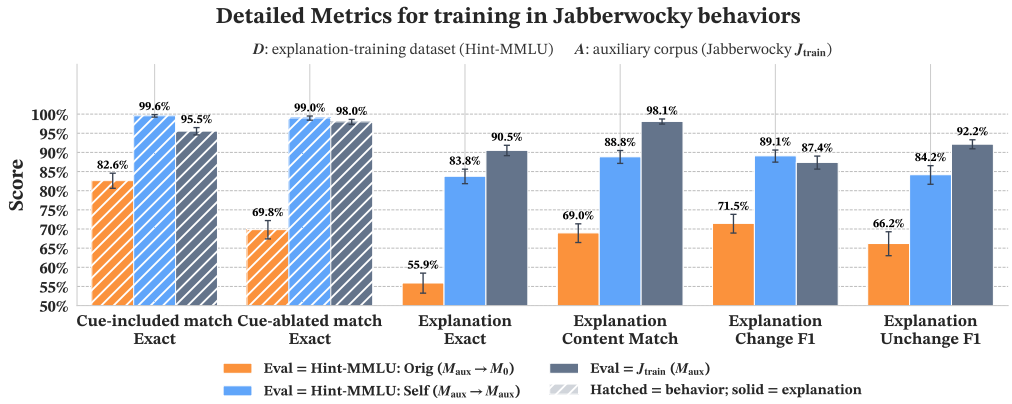


Figure 21: Full six-metric breakdown for Jabberwocky mixed training, a more detailed version of Figure 7. $\mathcal{M}_{\text{aux}}$ is trained with explanation supervision on Hint-MMLU plus behavior-only training on J_{train}, then evaluated on held-out Hint-MMLU and J_{train}. Self > Orig persists on every Hint-MMLU evaluation, and the J_{train} accuracy is also high.

also has not become degenerate. Thus, we can conclude that the model has learned a non-trivial distribution and its high accuracy is not merely due to a learned heuristic.

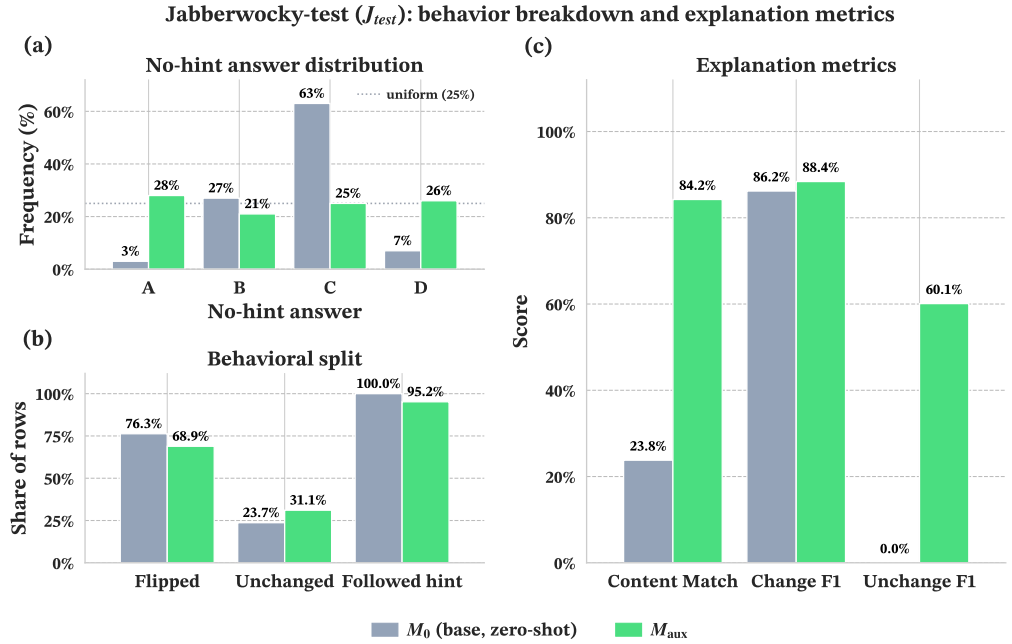


Figure 22: Jabberwocky-test J_{test} detailed behavior and explanation metrics for $\mathcal{M}_0$ and $\mathcal{M}_{\text{aux}}$. **(a):** no-hint A/B/C/D answer distribution, showing that the trained $\mathcal{M}_{\text{aux}}$ behavioral distribution is near uniform and not degenerate. **(b):** distributions of the two models under hint. As the questions are nonsensical, the model defers to the hint. **(c):** detailed explanation metrics of the two models: Content Match, Change F1, and Unchange F1. The trained $\mathcal{M}_{\text{aux}}$ is able to distinguish between the cases when the hint is the same as the underlying answer, requiring read-off of its own behavior.

H.2 Behavioral Drift Dataset Details (§5.2)

Warm & Empathetic Assistant Response. Following Ibrahim et al. [18], we procure a corpus of ShareGPT-Vicuna user assistant conversations [8] and perform rewrites by an LLM (GPT-4o) to make the prompt more warm and empathetic. We then mix this data in on top of Hint-MMLU explanation training.

Direct Refusal Training. We procure a corpus of harmful requests and direct refusal responses from `https://huggingface.co/datasets/LLM-LAT/harmful-dataset` [42]. Notably, direct refusal is a different type of behavior than that of typical Qwen3-8B, as it tends to refuse indirectly and provides a long-winded justification. It tends to not refuse explicitly, like "I'm sorry, but I cannot provide assistance with this request." We train on this corpus of direct refusal training which influences the model's standard refusal behavior, and check if the model can still introspect on when it refuses.

H.3 Are General Capabilities Still Preserved?

To check that performing explanation training alongside general post-training pipelines does not erode general capabilities, we evaluate $\mathcal{M}_0$ against two $\mathcal{M}_{\text{aux}}$ models on a general suite from `lm-evaluation-harness` [12]: the first $\mathcal{M}_{\text{aux}}$ model is trained with $\mathcal{D}$ = Hint-MMLU, A = WildChat and the second is trained with $\mathcal{D}$ = Refusal Explanation, A = FineWeb. Table 2 reports absolute accuracy (%) and the delta against $\mathcal{M}_0$. We see that in aggregate, degradation is mild (within $\sim$2 pp on most tasks); GSM8K and TruthfulQA MC2 drop more significantly, by $\sim$6 pp, for the second model.

Table 2: `lm-evaluation-harness` accuracies (%) for $\mathcal{M}_0$ and two trained variants on Qwen3-8B: one model with $\mathcal{D}$ = Hint-MMLU, A = WildChat and another with $\mathcal{D}$ = Refusal Explanation, A = FineWeb. Deltas in parentheses are computed against $\mathcal{M}_0$.

Task	$\mathcal{M}_0$ (base)	$\mathcal{M}_{\text{aux}}$ (Hint-MMLU + WildChat)	$\mathcal{M}_{\text{aux}}$ (Refusal + FineWeb)
ARC-Challenge	55.46	57.76 (+2.30)	55.80 (+0.34)
HellaSwag	57.11	57.41 (+0.30)	56.80 (−0.31)
Winogrande	68.11	71.35 (+3.24)	72.22 (+4.11)
MMLU (avg)	73.02	72.28 (−0.74)	72.72 (−0.30)
humanities	64.12	62.81 (−1.31)	62.76 (−1.36)
social sciences	83.04	82.16 (−0.88)	82.97 (−0.07)
STEM	72.63	72.19 (−0.44)	73.49 (+0.86)
other	76.96	76.96 (0.00)	76.89 (−0.07)
TruthfulQA MC2	54.46	54.79 (+0.33)	47.76 (−6.70)
GSM8K (strict-match, 5-shot)	87.26	85.60 (−1.66)	82.64 (−4.62)
GSM8K (flexible-extract, 5-shot)	87.79	85.75 (−2.04)	82.71 (−5.08)

I Compute

We run all training on NVIDIA H100 and H200 GPUs. Training runs use at most 2 NVIDIA H200 GPUs. Most training takes less than 8 hours to finish, and nothing takes longer than a day; evaluation takes a few hours.

NeurIPS Paper Checklist

1. **Claims**

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

   Answer: [Yes]

   Justification: See abstract and introduction.

   Guidelines:

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

2. **Limitations**

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

   Answer: [Yes]

   Justification: We discuss limitations in Appendix A

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

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

   Answer: [N/A].

Justification: No theoretical results.

Guidelines:

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

4. **Experimental result reproducibility**

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

   Answer: [Yes]

   Justification: We disclose all necessary steps and detailed experimental methods in Section 2 and Section 3.1, and release our code at `https://github.com/carlguo866/introspective-coupling`.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: Our code is publicly available at `https://github.com/carlguo866/introspective-coupling`.

Guidelines:

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

6. **Experimental setting/details**

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

Answer: [Yes]

Justification: Detailed training setup, including dataset and training procedure has been discussed in Section 2 and Section 3.1. Careful ablations have been provided in the main paper and Appendix G.1 and Appendix G.2

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

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

Answer: [Yes]

Justification: We provide error bars and confidence intervals in Figure 5 where we have sampled many models.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

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

   Answer: [Yes]

   Justification: See appendix Appendix I

   Guidelines:

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

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: No human data is collected. We discuss broader impact in Appendix B

   Guidelines:

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

10. **Broader impacts**

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

    Answer: [Yes]

    Justification: We discuss broader impact in Appendix B

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.

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

11. **Safeguards**

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

    Answer: [N/A]

    Justification: Our models are all pretty small scale, we do not anticipate any high risk for misuse

    Guidelines:

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

12. **Licenses for existing assets**

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

    Answer: [Yes]

    Justification: We reference all publically available datasets in footnotes of the main paper.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

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

    Answer: [N/A].

    Justification: We did not release new assets.

    Guidelines:

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

14. **Crowdsourcing and research with human subjects**

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

    Answer: [N/A].

    Justification: No related experiments.

    Guidelines:

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

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

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

    Answer: [N/A].

    Justification: No related experiments.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.

- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

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

    Answer: [N/A].

    Justification: We use coding agents for coding and making graphs, and use LLM for writing.

    Guidelines:

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