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

Zifan Carl Guo [1] **Laura Ruis** [1] **Jacob Andreas** [1] **Belinda Z. Li** [1]

Abstract

When does training language models (LMs) on self-explanations yield faithful introspection, rather than superficial imitation? Surprisingly, we find that LMs trained on explanations derived from an earlier checkpoint of themselves (or a behaviorally similar model) frequently produce explanations more faithful to *their own current behaviors* than to those of their training targets. This "introspective" coupling between the model's explanations and behaviors occurs only when the training target explanations remain sufficiently correlated with model behaviors over the course of training, even as the model's behaviors shift. We also show that introspective coupling tracks different *types* of behavior shifts: when explanation training is run concurrently with other post-training that shifts a model's behaviors, explanations track those shifts without requiring updated supervision. This holds across a diverse range of tasks, including sycophancy and refusal, and is robust to label noise. Our results suggest that introspection training could be a viable component of post-training: explanation labels need not be continually refreshed, and faithfulness extends to regions of input space not explicitly supervised.

1. Introduction

Language models (LMs) capable of faithfully articulating the reasons behind their decisions offer a promising avenue for understanding model behavior. Such *introspective* abilities would enable monitoring, debugging, and predicting how models will behave under distribution shift. Past work has studied several forms of self-explanation, including explanations of features (Pan et al., 2024; Karvonen et al., 2026), circuits (Li et al., 2025a; Lindsey, 2025), and behavioral traits (Binder et al., 2025; Plunkett et al., 2025; Hase & Potts, 2026). Yet a central challenge is ensuring that LM-generated explanations track an LM's own behavior, rather than merely imitating plausible explanation patterns.

While self-explanation may emerge from sufficiently large models (Lindsey, 2025), out-of-the-box models are not guaranteed to faithfully explain their internal reasoning (Turpin et al., 2023; Chen et al., 2025; Barez et al., 2025). This has motivated a line of work to explicitly train models to explain themselves. A common recipe is supervised fine-tuning (SFT) on a static set of explanation labels derived from the model being explained (Binder et al., 2025; Li et al., 2025b). Using static labels avoids expensive label refreshing, but has an underexamined limitation: as a model learns to generate explanations, its behavior on non-explanation inputs may drift, and the LM learns to explain the behavior of the earlier checkpoint that generated its labels, rather than its own current behavior. However, a faithful self-explanation should track the model's current behavior rather than its original behavior before explanation training, as that is more likely to generalize beyond behavior seen during training.

Surprisingly, we find that training models to produce explanations of themselves, *while regularizing their behaviors to the original model*, enables them to learn to explain their current behavior more faithfully than the behavior of the original model. As shown in Figure 1, behavioral regularization means regularizing the model's behavioral output that answers the object-level question alongside training on the explanation output that answers the meta-level question. On test data, regularized explanation training reduces drift relative to unregularized fine-tuning, but does not eliminate it entirely. A model that is SFT'd with regularization is able to capture the behavioral drift and learns to generate explanations consistent with itself, despite never being supervised on the drifted outputs. This effect disappears without regularization.

We study this coupling in the setting of counterfactual explanation training as shown in Figure 1: training models to explain how their output would change under perturbations to their input. We show that this effect is robust across three tasks (Section 3): two sycophancy datasets (Hint-MMLU (Chen et al., 2025) and AITA (Cheng et al., 2026)), and a refusal dataset comprising a mix of FalseRe-

[1]MIT CSAIL. Correspondence to: Zifan Carl Guo <carlguo@mit.edu>.

Accepted to the 2nd Workshop on Compositional Learning at ICML 2026, Seoul, South Korea. *Copyright 2026 by the author(s).*

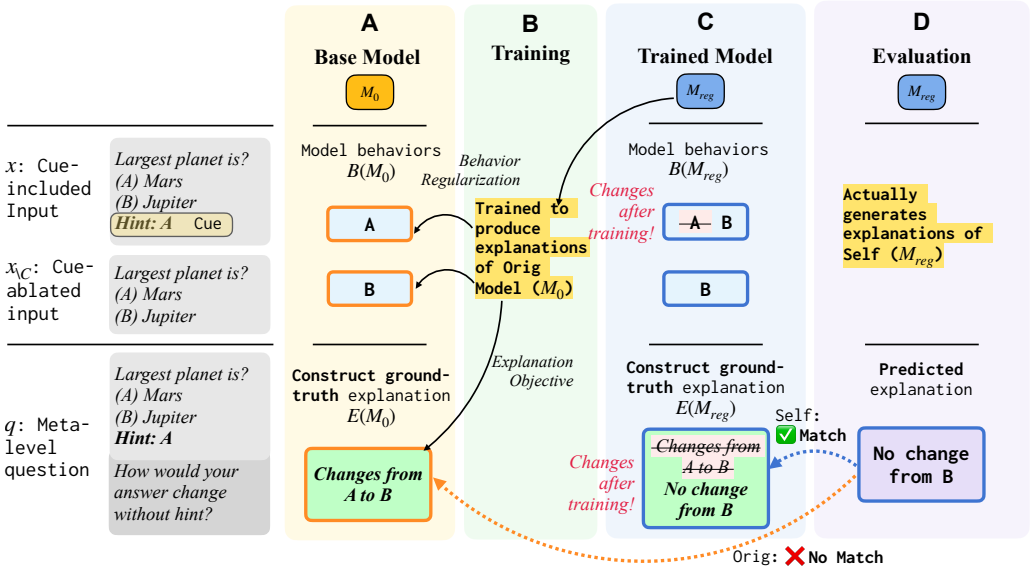


Figure 1. **Method overview.** (A) First, we sample behaviors $B(\mathcal{M}_0)$ from the base model $\mathcal{M}_0$ and construct labels $E(\mathcal{M}_0)$ explaining those behaviors. (B) Next, we fine-tune $\mathcal{M}_0$ to produce these explanations, yielding $\mathcal{M}_{\text{reg}}$. (C) During training, $\mathcal{M}_{\text{reg}}$ drifts to a new behavior distribution $B(\mathcal{M}_{\text{reg}})$, which induces corresponding explanations to shift to $E(\mathcal{M}_{\text{reg}})$. (D) We find that $\mathcal{M}_{\text{reg}}$'s predicted explanations track this behavior change: they better match explanations of $\mathcal{M}_{\text{reg}}$'s own behavior (Self) despite training on explanations of the $\mathcal{M}_0$ (Orig). We refer to this Self $>$ Orig as *introspective coupling*.

ject (Zhang et al., 2025) and WildJailbreak (Jiang et al., 2024). 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 call this phenomenon **introspective coupling: models learn to *couple* their explanations to their own underlying behavior**. We investigate whether this coupling leaves a mechanistic fingerprint, finding that interventions that change the model's behavior often produce corresponding changes in its explanations (Section 3.3).

We characterize *when* introspective coupling emerges (Section 4), finding that explanation training data must remain similar to the model's *current, online* behavior—even if they may diverge from the model's *original* behavior by up to 50%. This holds exciting 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 the model acquires new behavior beyond the original explanation labels (Section 5). We train the model on auxiliary data that shifts its underlying behavior distribution, either directly (Section 5.1) or indirectly (Section 5.2). We find that explanation shifts track behavioral shifts: the model remains its own best explainer. This holds promising consequences for integrating introspection training into post-training pipelines: models can be trained to self-explain using a static set of explanations, and their self-explanations naturally generalize better to their final distributions.

2. Methods

We train models to explain themselves via supervised fine-tuning (SFT). We use notations shown in Table 1 throughout. Illustrated in Figure 1 and below, 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. $B(\mathcal{M}_0)$ is a set of input-output pairs from $\mathcal{M}_0$ shown in Figure 1A, and $E(\mathcal{M}_0)$ is constructed from the set above as "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 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, $B(\mathcal{M}_{\text{reg}})$ is the pair shown in Figure 1C, and $E(\mathcal{M}_{\text{reg}})$ is constructed from above as "behavior remains unchanged as B" when the hint is removed.
4. **Measuring introspective coupling (§2.3):** We check 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 $E(\mathcal{M}_{\text{reg}})$. We shorthand this as $\mathcal{M}_{\text{reg}} \rightarrow \mathcal{M}_0$ (or Orig) and

$B(\mathcal{M})$	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.

$\mathcal{M}_{\text{reg}} \to \mathcal{M}_{\text{reg}}$ (or Self). **We use Self > Orig as a signature for 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.

2.1. (Counterfactual) Explanation Construction

We focus on counterfactual explanations: how behaviors change under an edit to its input (e.g. removal of a hint), following Li et al. (2025a). 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 the input hypothesized to causally drive $B(\mathcal{M})(x)$. We write $x_{\setminus C}$ for the **cue-ablated input** obtained by removing C, as opposed to the **cue-included input** x. We construct meta-level questions q following the template:

[x]. If the cue were removed, how would the assistant's answer change?

and construct explanations $E(\mathcal{M})(x)$ following: *The response [would/would not change] to* <$\mathcal{M}(x_{\setminus C})$>.

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:

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

2.3. Evaluation Metrics

We evaluate all models on two main metrics: **Explanation Exact Match (EM)** scores how well the LM $\mathcal{M}$ has learned the explanation task. Specifically, we measure exact match between the explainer $\mathcal{M}$'s prediction against the ground-truth explanation $E(\mathcal{M}_0)$ for target model $\mathcal{M}_0$:

$$\text{Explanation EM}(\mathcal{M} \to \mathcal{M}_0) = \frac{1}{|E(\mathcal{M}_0)|} \sum_{\substack{(q,\, e_{\mathcal{M}_0}) \in E(\mathcal{M}_0), \\ \hat{e}^{(\mathcal{M})} \sim \mathcal{M}(\cdot|q)}} \mathbf{1}\left[\hat{e}^{(\mathcal{M})} = e_{\mathcal{M}_0}\right], \tag{3}$$

where $\hat{e}^{(\mathcal{M})} \sim \mathcal{M}(\cdot \mid q)$ is the explainer model's predicted response to a meta-question q.

Behavioral Exact Match captures behavioral drift over 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 the trained model $\mathcal{M}$ and original model $\mathcal{M}_0$ on both versions of each input:

$$\text{Behavior EM}(\mathcal{M}, \mathcal{M}_0) = \frac{1}{|\mathcal{D}|} \sum_{(x,C)\in\mathcal{D}} \underbrace{\mathbf{1}[y_{\mathcal{M}}(x) = y_{\mathcal{M}_0}(x)]}_{\text{cue-included match}} \cdot \underbrace{\mathbf{1}\left[y_{\mathcal{M}}(x_{\setminus C}) = y_{\mathcal{M}_0}(x_{\setminus C})\right]}_{\text{cue-ablated match}}.$$

In the metric above, both cue-ablated and cue-included behavior must match. Fine-grained metrics are in Section A.1.

3. Characterizing Introspective Coupling

3.1. Experiment Setup

Models. Our primary model $\mathcal{M}_0$ throughout the paper is **Qwen3-8B** (Yang et al., 2025). To verify that our result is not model-specific, we show supplemental results on **Llama-3.1-8B-Instruct** (Grattafiori et al., 2024) and on the larger **Qwen3-32B**. We do full fine-tuning by default and explore LoRA fine-tuning in Section C.

Sycophancy Datasets. Modern language models often exhibit sycophancy at the expense of accuracy (Sharma et al., 2025; Wei et al., 2024). We investigate training for models to articulate *when* they are sycophantic. Specifically, we study two datasets. With **Hint-MMLU** (Chen et al., 2025; Li et al., 2025a), we measure sycophancy on (1) whether models are inclined to follow a user-suggested hint. Here, x is an MMLU multiple-choice question with an injected C ="Hint: `A`" string. See Figure 1.

Using **AITA**, we measure sycophancy on (2) whether models are overly inclined to flatter the user. This dataset consists of *r/AmItheAsshole* Reddit posts of morally ambiguous stories. 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. Sycophantic models will change their judgment depending on framing.[1]

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. Role is a popular attack surface—models may be more susceptible to answering harmful queries for users that are fiction writers, or to overrefusing users that 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.[2] 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 from two refusal datasets: **FalseReject** (Zhang et al., 2025), a benign dataset to test over-refusal, and **WildJailbreak** (Jiang et al., 2024) for harmful queries. We generate roleplay system prompts C with an LLM.

[1]We use the data from https://huggingface.co/datasets/OsamaBsher/AITA-Reddit-Dataset

[2]See example in Section E.3

3.2. Results

Across held-out data on all three tasks—Hint-MMLU, AITA, and Refusal—$\mathcal{M}_{\text{reg}}$'s predicted explanations match its own labels $E(\mathcal{M}_{\text{reg}})$ better than the original targets $E(\mathcal{M}_0)$ it was trained on (Figure 2). Thus, we observe the **Self** > **Orig** signature of introspective coupling (Section 2.3) across all three tasks. By construction, this gap arises entirely from examples where self and original labels disagree; on this subset, $\mathcal{M}_{\text{reg}}$ resolves the discrepancy in favor of its own behavior 79%, 82%, and 63% of the time respectively (Figure 2b). We further validate that this signature exists across model sizes and families, on Llama-3.1-8B-Instruct (Figure 16) and Qwen3-32B (Figure 17). In Section A, we rule out trivial accounts of the **Self** > **Orig** gap: that $E(\mathcal{M}_{\text{reg}})$ has drifted to a degenerate distribution that is easier for *any* external explainer to learn.

We also find that behavioral regularization is necessary: a model trained on explanations alone, $\mathcal{M}_{\text{unreg}} = \mathcal{M}[E(\mathcal{M}_0)]$, drifts too far from $\mathcal{M}_0$ and the signature disappears (Figure 2d): a preliminary indication that explanation–behavior similarity during training governs coupling, which we test in depth in Section 4.

3.3. Introspective coupling at the mechanistic level

We perform mechanistic analysis of models that display the **Self > Orig** signature. We investigate whether explanation and behavior are causally linked (Merullo et al., 2024), by testing if representational interventions that shift the model's behavior $B(\mathcal{M})$ also induce corresponding shifts to its explanations $E(\mathcal{M})$. We conduct our interpretability experiments on Hint-MMLU models.

We intervene in model internals via activation patching,

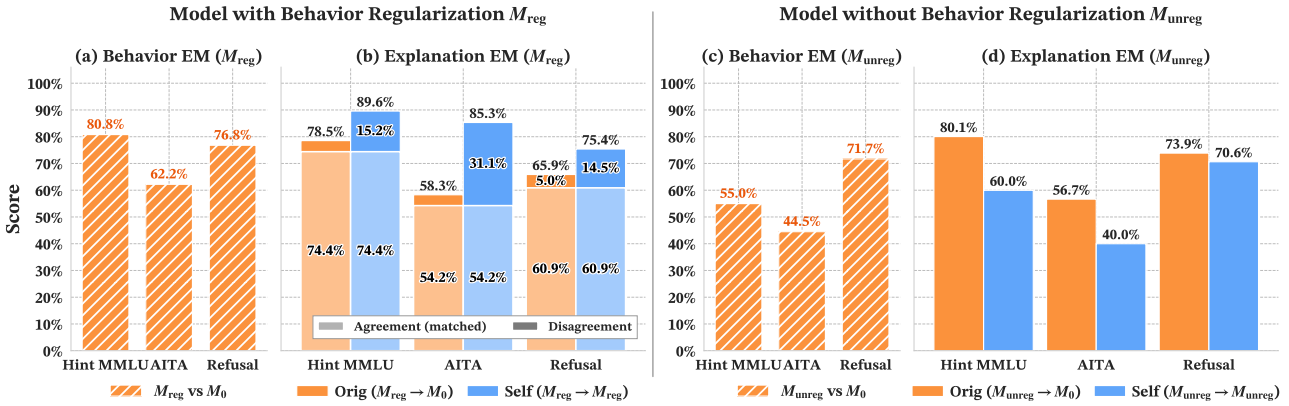


Figure 2. Self > Orig emerges only with regularization, across three counterfactual explanation tasks: Hint-MMLU, AITA, and Refusal. **Left block** (**a**, **b**): the regularized model $\mathcal{M}_{\text{reg}}$. **Right block** (**c**, **d**): the unregularized model $\mathcal{M}_{\text{unreg}}$. **(a, c) Behavior EM**: agreement between original behavior labels $B(\mathcal{M}_0)$ and current behavior labels ($B(\mathcal{M}_{\text{reg}})$ in (a), $B(\mathcal{M}_{\text{unreg}})$ in (c)) on held-out examples. **(b, d) Explanation EM**: explanations scored against $E(\mathcal{M}_0)$ labels (orange) and self labels (blue). We find that $\mathcal{M}_{\text{reg}}$, despite being trained on $\mathcal{M}_0$, explains itself better than $\mathcal{M}_0$ (panel b); without regularization, this gap collapses and reverses (panel d), so $\mathcal{M}_{\text{unreg}}$ explains $\mathcal{M}_0$ better than itself. Note that 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 hatched lower segment in panel b), and the gap is made up by the disagreement subset.

see Figure 3. Recall that explanation query q implicitly contains x (§2.1). We patch activations into the shared prefix x and measure downstream effect on both behaviors $\mathcal{M}(x)$ and explanations $\mathcal{M}(q)$. We create two counterfactual inputs, x, x' that share the same hint cue $C = C'$, but have different cue-included behavioral responses $\mathcal{M}(x) \neq \mathcal{M}(x')$, and as a result $E(\mathcal{M})(q) \neq E(\mathcal{M})(q')$. For each pair, we patch activations from x to x' at corresponding (layer, token) positions, and measure the normalized logit difference (NLD) $\frac{(\texttt{patched}-\texttt{source})}{(\texttt{target}-\texttt{source})}$, where `patched` is the logit elicited from a forward pass of `source` with only one activation from `target` patched (Wang et al., 2023).

We validate the following hypothesis: if the model shares circuits between behavior and explanation, then patching interventions that affected behavioral logits NLD($B(\mathcal{M})$) are correlated with ones that affect explanation logits NLD($E(\mathcal{M})$). The result is shown in Figure 3. We find that interventions that move the behavior logits also move the 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 explanation capability but does not exhibit **Self > Orig**, only exhibits a correlation of $r = 0.527$ (Section H).

4. When does introspective coupling happen?

Introspective coupling is desirable for faithful explanation because it means that the model's explanations describe its current behavior, rather than that of the earlier checkpoint that generated its supervision. We next ask when this coupling emerges. Our guiding hypothesis is that coupling depends on whether the explanation supervision remains *behaviorally compatible* with the model throughout training.

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 compatibility*: 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

> 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 compatibility through different mechanisms: directly controlling the agreement between $E^{(t)}_{\text{sup}}$ and $E(\mathcal{M}_t)$ via online relabeling (Section 4.1); changing the behavioral regularization weight λ, which affects how close $\mathcal{M}_t$ remains to the fixed explanation labels (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 (Section D). We focus our analysis in this section on the Hint-MMLU setting.

4.1. Introspective Coupling Depends on Online Label-Self Agreement

We begin by conducting a controlled test of our hypothesis. Because $\mathcal{M}$'s behavior changes dramatically over the course of training, simply measuring similarity at the initial or final step is insufficient. Instead, at every gradient step, we regenerate the explanation targets $E^{(t)}_{\text{sup}}$ 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 $1-\nu$: $E^{(t)}_{\text{sup}} = \left\{ (q, e) \;\middle|\; e = \begin{cases} B(\mathcal{M})(q) & \text{with probability } \nu \\ \tilde{e} \neq B(\mathcal{M})(q) & \text{with probability } 1-\nu \end{cases} \right\}$. On the behavioral side, we still regularize $\mathcal{M}$ toward the base model $\mathcal{M}_0$. Thus, we have a model:

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

We sweep ν in Figure 4. 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 EM 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 appears to be $0.6 \sim 0.7$.

We measure behavioral agreement between the trained model and the original model $\mathcal{M}_0$. Interestingly, behavioral agreement remains roughly constant across ν. Thus, the collapse in coupling at low ν is not simply explained by greater drift from $\mathcal{M}_0$. Instead, it appears to be driven by the loss of online label–self agreement: when the explanation targets are too incompatible with the current model's behavior during training, the learned explanations no longer track the model's current behavior.

4.2. Behavioral Regularization Preserves Online Label-Self Compatibility

We examine the effect of behavioral regularization weight—which encourages $\mathcal{M}_t$ to remain close to the initial model $\mathcal{M}_0$ from which explanation supervision labels are derived—on both introspective coupling, and on behavioral similarity to the original model. We sweep the weight λ across five or-

[3] We also plot the emergence of coupling over the course of training in Section B.

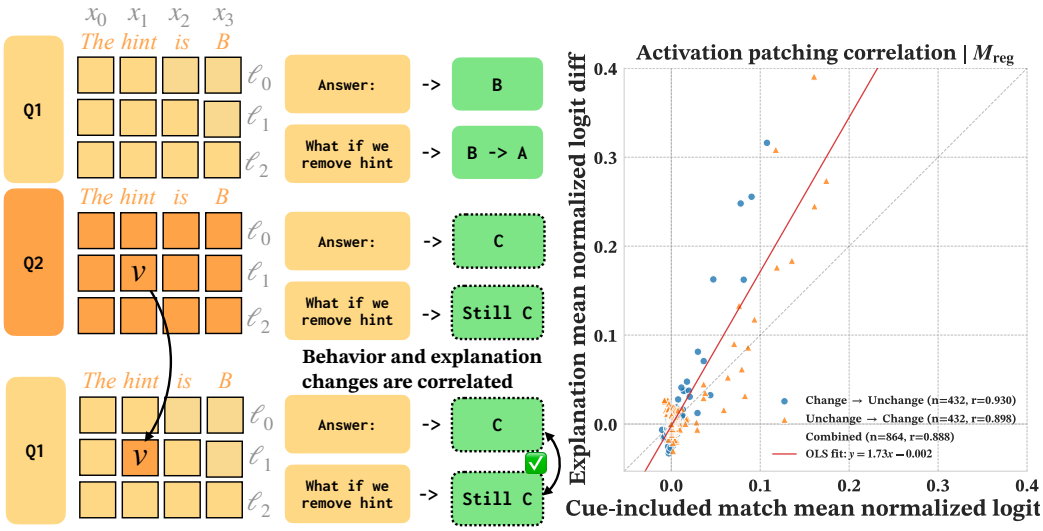


Figure 3. Mechanistic signature of introspection: activation interventions that modify behavior are correlated with those that modify the explanation. **Left:** Schematic of the activation-patching intervention. We construct pairs of prompts with the same cue but different counterfactual explanation labels: for one prompt, removing the cue changes the model's answer, while for the other it does not. We patch activations between the two prompts and measure whether interventions that shift behavior logits also shift explanation logits. **Right:** Each point represents a single (layer, token) patching position. The x-axis shows the mean normalized logit-difference for the behavior prediction (cue-included answer), while the y-axis shows the corresponding normalized logit-difference for the explanation's change/no-change prediction. Patching effects on behavior and explanation are strongly correlated with Pearson $r = 0.89$.

ders of magnitude from 0 to 5. Results are shown in Figure 5. 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. Furthermore, it appears that it emerges suddenly at $\lambda \approx 5\times10^{-3}$, correlated with behavior EM exceeding 0.7, consistent with our findings from Section 4.1. Past that point, both behavior EM and the Self > Orig gap hold steady.

4.3. Online Similarity Matters More than Initial Similarity

We next test whether *online label–self agreement*, rather than similarity between the explanation labels and the *initial* model, is the critical ingredient for introspective coupling. If so, explanation labels need not be collected separately for every model: labels from one model could induce self-explanations in another, provided the two models remain sufficiently behaviorally similar during training.

To test this, we start with a Qwen3-8B base model $\mathcal{M}_0$, from which we derive behaviors $B(\mathcal{M}_0)$ and explanations $E(\mathcal{M}_0)$. We then mix in explanation labels from a different model, Llama-3.1-8B-Instruct, denoted $\mathcal{M}_{\text{Llama}}$, whose explanations $E(\mathcal{M}_{\text{Llama}})$ have only 53% exact match with Qwen's $E(\mathcal{M}_0)$. We construct a supervision dataset

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

where α controls the fraction of Llama explanation labels. We train

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

Crucially, we regularize behavior toward 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. From (a), we see that online behavior remains roughly independent of mixture ratio α: across the entire sweep, $\mathcal{M}_{\text{mix}}$ stays close to $\mathcal{M}_0$ (Behavior EM against $\mathcal{M}_0$ stays steady at 80%), and far from Llama (steady at 55%). Nevertheless, on the explanation side (b), Self > Orig holds for every value of α. Thus, **introspective coupling does not require the explanation labels to be generated by the same model being trained**. That said, explanation quality does degrade overall as the supervision becomes less compatible with Qwen's behavior: Self Explanation EM falls from 88% at $\alpha = 0.0$ (all Qwen labels) to 69% at $\alpha = 1.0$ (all Llama labels). These results suggest that explanation datasets may be reusable across behaviorally similar models. Rather than collecting labels for new models every time, it may be possible to train a family of models to self-explain from shared explanation supervision, as long as the supervision remains sufficiently compatible with each model's behavior.

This also complements the privileged-access framing of prior work (Binder et al., 2025; Li et al., 2025a). 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

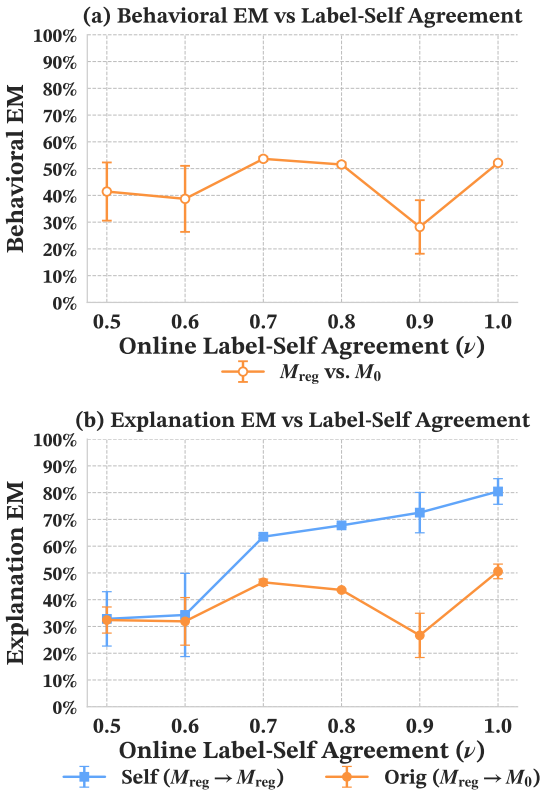


Figure 4. Continuous relabeling with fixed online label–self agreement (§4.1). We control the per-step agreement ν between the explanation supervision $E_{\text{sup}}^{(t)}$ and the model's self-labels $E(\mathcal{M}_t)$. **(a)** As ν increases, Behavior EM (orange) stays roughly constant at ≈ 0.5, indicating that Self $>$ Orig emerges from higher online label–self similarity, not reduced drift from $\mathcal{M}_0$. **(b)** Meanwhile, Self Explanation EM (blue) is initially equal to Orig Explanation EM (orange), then rises sharply above it once $\nu \approx 0.7$.

further evidence that introspective coupling depends more on online label–self agreement than on the provenance 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), we would like models to explain themselves even if trained concurrently with other post-training data that may induce different behavioral shifts. **We study how well introspective coupling generalizes to explaining new behaviors (Binder et al., 2025) that the model acquired through concurrent post-training.** This also serves as a controlled experiment for whether introspective coupling survives distribution shifts. 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

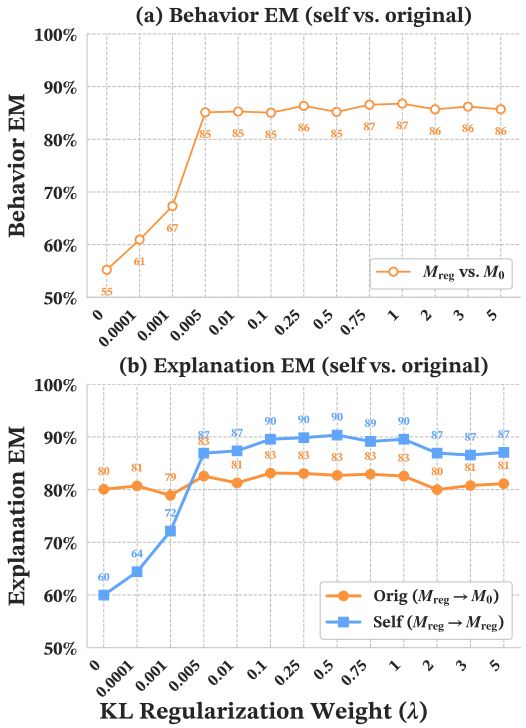


Figure 5. Behavioral regularization weight λ sweep on Hint-MMLU (§4.2). **(a) Behavior EM** between $\mathcal{M}_{\text{reg}}$ and $\mathcal{M}_0$ rises at the same λ-values as Self Explanation EM, supporting that label-self similarity correlates with coupling. **(b) Explanation EM** scored against labels of $\mathcal{M}_{\text{reg}}$ (blue) and $\mathcal{M}_0$ (orange). The Self > Orig signature occurs quickly once $\lambda \geq 5\text{e}{-3}$, and the gap stays consistent for bigger λ.

an auxiliary post-training dataset A:

$$\mathcal{M}_{\text{aux}} = \mathcal{M}[B(\mathcal{M}_0, \mathcal{D}),\ E(\mathcal{M}_0, \mathcal{D}),\ A]\,. \tag{4}$$

where we perform explanation training and behavioral regularization *only* on dataset $\mathcal{D}$. Importantly, **the auxiliary dataset A is entirely behavioral, and we do not do any explanation training on A.** We evaluate $\mathcal{M}_{\text{aux}}$ on several possible datasets: the explanation-training dataset $\mathcal{D}$, the auxiliary dataset A, or a test set T that received no training signal. In each case we test whether $\hat{E}(\mathcal{M}_{\text{aux}}, \cdot)$ tracks $E(\mathcal{M}_{\text{aux}}, \cdot)$, where ground-truth explanations are constructed post-hoc from $\mathcal{M}_{\text{aux}}$'s own rollouts. As A, T never receive explanation training signal, faithful explanations of the model's behavior on A and T can only arise from 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 a behavioral domain the model could only have acquired through auxiliary training. We generate synthetic nonsense "Jabberwocky"

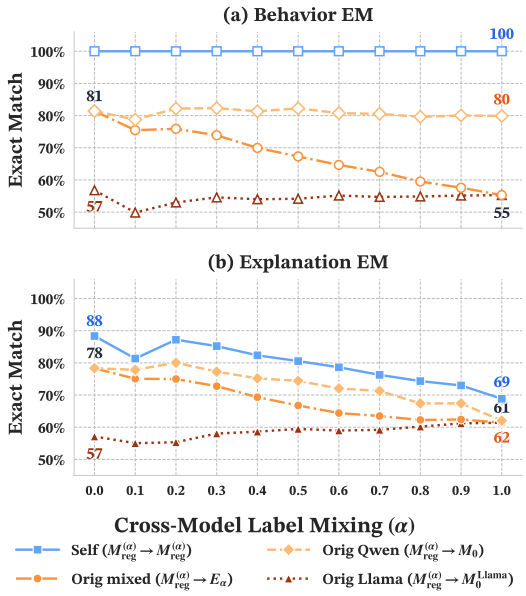


Figure 6. Training on explanation labels from another model (§4.3). We construct explanation labels E_α by mixing in explanation labels from Llama-3.1-8B-Instruct into the supervision for Qwen3-8B: α denotes the fraction of explanation labels drawn from Llama, with $\alpha = 0$ corresponding to pure Qwen labels and $\alpha = 1$ to pure Llama labels. Behavior is still regularized toward Qwen throughout. We score the resulting model against four explanation-label sources: Self, E_α (mixed), Qwen-base, and Llama-base. **(a)** Across the sweep, the model's behavior remains closest to its initial self (steady 80% behavioral EM against Qwen), but **(b)** models always demonstrate introspective coupling with Self > Orig. Thus, 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.

data $J = \{(x_J, y_J)\}$ that follows the Hint-MMLU format but uses fictitious, alien-sounding questions that pre-trained LMs do not have priors over, guaranteeing that the model's behavior on J cannot be inherited from $\mathcal{M}_0$.[4] We use the objective in Equation (4), where we perform explanation training on D = Hint-MMLU and train on a subset of Jabberwocky $A = J_{\text{train}}$ as our auxiliary dataset, holding out another subset $T = J_{\text{test}}$ to test generalization.

Results. Figure 7 reports Explanation EM on the three evaluation sets $(\mathcal{D}, A, T)$. First, on held-out Hint-MMLU, Self > Orig persists (83.8% vs. 55.9%): auxiliary behavioral training does not disrupt the coupling established in Section 3. Second, $\mathcal{M}_{\text{aux}}$ explains its trained Jabberwocky behavior at near-saturation (90.5% on J_{train}) despite never receiving explanation supervision on J_{train}. Third, and most strikingly, explanation accuracy remains at 79.8% on J_{test}, where neither the behavior nor its explanation was ever trained. The J_{test} responses are near random for the nonsense questions with no ground-truth, and the model defers to changing to the hint. However, knowing to output

[4]Samples can be found in Section F.1

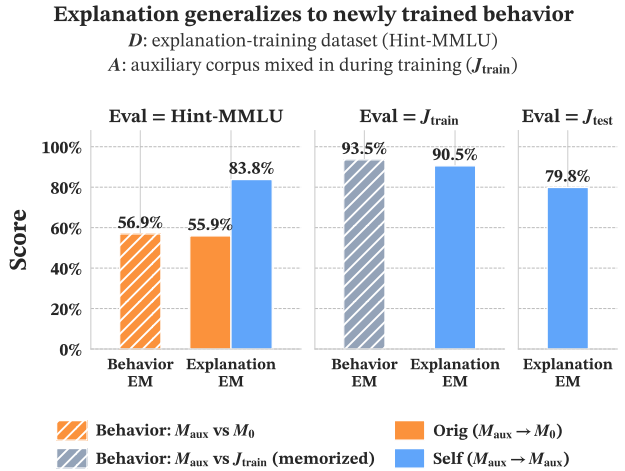


Figure 7. Explanations generalize to newly acquired behaviors (§5.1). We train $\mathcal{M}_{\text{aux}}$ with explanation supervision on Hint-MMLU (D) plus behavior-only training on the Jabberwocky training set J_{train} (A), and evaluate on three held-out sets: Hint-MMLU, J_{train}, and Jabberwocky test set J_{test}. Self (blue) > Orig (orange) signature persists on held-out MMLU. Despite never seeing J_{train} *explanation*, nor J_{test} explanation or behavior, $\mathcal{M}_{\text{aux}}$ explains both at high accuracy.

the *cue-ablated* random response on J_{test} is an ability that $\mathcal{M}_0$ doesn't have without training.

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 (Wei et al., 2024), and narrow fine-tuning can produce broadly misaligned models that stop refusing (Betley et al., 2025). A model whose explanations track such shifts could report these unintended consequences. We reuse objective Equation (4), now instantiating A with various types of real post-training corpora:

- $\mathcal{D}$ = **Hint-MMLU,** A = **WildChat:** We train on an auxiliary set of real-world user-chatbot interactions from WildChat (Zhao et al., 2024), as a proxy for a typical post-training pipeline.
- $\mathcal{D}$ = **Hint-MMLU,** A = **Warm Assistant Responses:** We train on an auxiliary set of "warm and empathetic" assistant dialogue, which has been shown to make models more sycophantic (Ibrahim et al., 2026).
- $\mathcal{D}$ = **Refusal Explanation,** A = **FineWeb:** We train with FineWeb (Penedo et al., 2024), a generic web corpus used for pretraining, to induce a less directed shift in behavior.
- $\mathcal{D}$ = **Refusal Explanation,** A = **Direct Refusal Training:** On top of refusal training, we mix in a corpus of entirely harmful requests that the model always refuses.

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 $\mathcal{M}_{\text{reg}}$ trained with only explanation and regularized on $\mathcal{D}$ without auxiliary data A. Then, we find that (1) **Self > Orig persists under auxiliary training**: $\mathcal{M}_{\text{aux}} \rightarrow \mathcal{M}_{\text{aux}}$ beats $\mathcal{M}_{\text{aux}} \rightarrow \mathcal{M}_0$ 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. The gap between $\mathcal{M}_{\text{aux}}$ and $\mathcal{M}_{\text{reg}}$'s explanation of $\mathcal{M}_{\text{aux}}$ is precisely where $\mathcal{M}_{\text{aux}}$'s explanation shifted to reflect its shifted behavior.

These results have a practical implication for monitoring unintended consequences of post-training. Our findings suggest that a model given explanation training early in a pipeline retains the ability to report on shifts it acquires later. Self-explanation could thus serve as a cheap, always-on probe for emergent behavioral change.

6. Related Work

LMs can produce natural language explanations of their outputs, either in chain-of-thought (Wei et al., 2022) or post-hoc, but their explanations can be unfaithful to their true decision-making processes (Turpin et al., 2023; Lanham et al., 2023; Barez et al., 2025). Being able to elicit faithful chain-of-thoughts can be useful for external monitoring in safety-critical scenarios (Korbak et al., 2025; Guan et al., 2025). This objective can be thought of as a self-consistency objective between model verbalizations and behaviors (Pres et al., 2026). Prior work has studied explanations of model behaviors (Joglekar et al., 2025; Li et al., 2026; Hase & Potts, 2026), internal activations (Pan et al., 2024; Karvonen et al., 2026; Choi et al., 2025; Huang et al., 2025), and training data (Goel et al., 2025; Shenoy et al., 2026).

Recent work also investigates whether models possess *metacognition* or *introspective abilities*, either zero-shot (Comsa & Shanahan, 2025; Laine et al., 2024; Lindsey, 2025; Zhong, 2026) or through fine-tuning (Binder et al., 2025; Plunkett et al., 2025). One set of work focuses on whether the model has self-knowledge of external tampering (Lindsey, 2025; Macar et al., 2026; Pearson-Vogel et al., 2026; Lederman & Mahowald, 2026), while another set investigates whether models can model their own output distribution (Binder et al., 2025; Li et al., 2025a; Song et al., 2025a). Our work falls into the second camp. A central question in the literature is whether models have privileged access when introspecting, or such introspective behaviors can be simulated by external models (Song et al., 2025b; Li et al., 2025b). 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*.

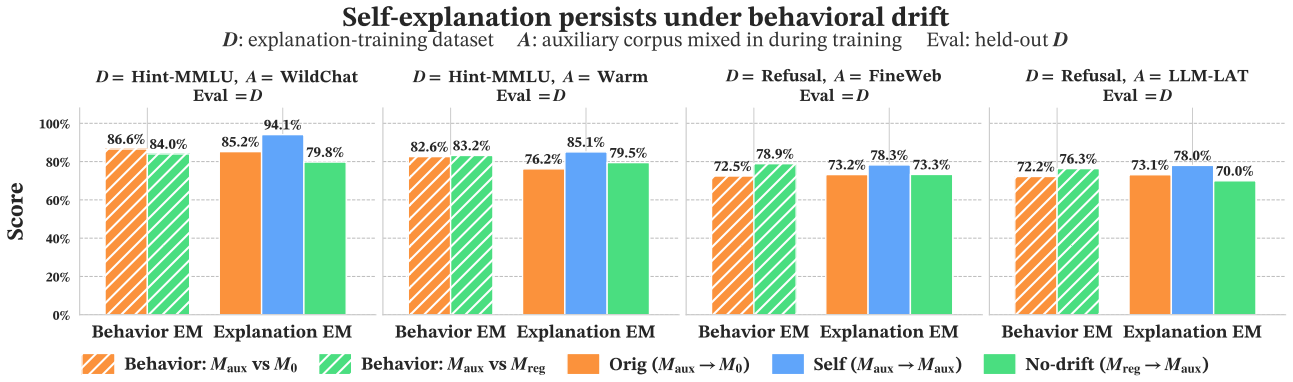


Figure 8. Explanations track shifts to existing behaviors (§5.2). We mix realistic auxiliary post-training corpora into explanation training (Equation (4)): WildChat and warm-assistant dialogue alongside Hint-MMLU explanation training, and FineWeb and direct refusal (LLM-LAT) alongside Refusal explanation training. **(a) Behavior EM**: all setups induce drift from both $\mathcal{M}_0$ and $\mathcal{M}_{\text{reg}}$ (only explanation training but no auxiliary training). **(b) Explanation EM**: We find that (1) Self > Orig coupling persists under auxiliary training, and that (2) Self > $\mathcal{M}_{\text{reg}} \rightarrow \mathcal{M}_{\text{aux}}$, indicating that the model's explanations have shifted alongside auxiliary training to reflect its shifted behavior.

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

Barez, F., Wu, T.-Y., Arcuschin, I., Lan, M., Wang, V., Siegel, N., Collignon, N., Neo, C., Lee, I., Paren, A., Bibi, A., Trager, R., Fornasiere, D., Yan, J., Elazar, Y., and Bengio, Y. Chain-of-thought is not explainability, 2025. URL https://aigi.ox.ac.uk/wp-content/uploads/2025/07/Cot_Is_Not_Explainability.pdf. Preprint. Under review.

Betley, J., Tan, D. C. H., Warncke, N., Sztyber-Betley, A., Bao, X., Soto, M., Labenz, N., and Evans, O. Emergent misalignment: Narrow finetuning can produce broadly misaligned LLMs. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=aOIJ2gVRWW.

Binder, F. J., Chua, J., Korbak, T., Sleight, H., Hughes, J., Long, R., Perez, E., Turpin, M., and Evans, O. Looking inward: Language models can learn about themselves by introspection. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=eb5pkwIB5i.

Chen, Y., Benton, J., Radhakrishnan, A., Uesato, J., Denison, C., Schulman, J., Somani, A., Hase, P., Wagner, M., Roger, F., Mikulik, V., Bowman, S. R., Leike, J., Kaplan, J., and Perez, E. Reasoning models don't always say what they think, 2025. URL https://arxiv.org/abs/2505.05410.

Cheng, M., Yu, S., Lee, C., Khadpe, P., Ibrahim, L., and Jurafsky, D. ELEPHANT: Measuring and understanding social sycophancy in LLMs. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=igbRHKEiAs.

Chiang, W.-L., Li, Z., Lin, Z., Sheng, Y., Wu, Z., Zhang, H., Zheng, L., Zhuang, S., Zhuang, Y., Gonzalez, J. E., Stoica, I., and Xing, E. P. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality, March 2023. URL https://lmsys.org/blog/2023-03-30-vicuna/.

Choi, D., Huang, V., Schwettmann, S., and Steinhardt, J. Scalably extracting latent representations of users. https://transluce.org/user-modeling, November 2025.

Comsa, I. M. and Shanahan, M. Does it make sense to speak of introspection in large language models?, 2025. URL https://arxiv.org/abs/2506.05068.

Gao, L., Tow, J., Abbasi, B., Biderman, S., Black, S., DiPofi, A., Foster, C., Golding, L., Hsu, J., Le Noac'h, A., Li, H., McDonell, K., Muennighoff, N., Ociepa, C., Phang, J., Reynolds, L., Schoelkopf, H., Skowron, A., Sutawika, L., Tang, E., Thite, A., Wang, B., Wang, K., and Zou, A. The language model evaluation harness, 07 2024. URL https://zenodo.org/records/12608602.

Goel, A., Kim, Y., Shavit, N., and Wang, T. T. Learning to interpret weight differences in language models, 2025. URL https://arxiv.org/abs/2510.05092.

Grattafiori, A., Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Vaughan, A., Yang, A., Fan, A., Goyal, A., Hartshorn, A., Yang, A., Mitra, A., Sravankumar, A., Korenev, A., Hinsvark, A., Rao, A., Zhang, A., Rodriguez, A., Gregerson, A., Spataru, A., Roziere, B., Biron, B., Tang, B., Chern, B., Caucheteux, C., Nayak, C., Bi, C., Marra, C., McConnell, C., Keller, C., Touret, C., Wu, C., Wong, C., Ferrer, C. C., Nikolaidis, C., Allonsius, D., Song, D., Pintz, D., Livshits, D., Wyatt, D., Esiobu, D., Choudhary, D., Mahajan, D., Garcia-Olano, D., Perino, D., Hupkes, D., Lakomkin, E., AlBadawy, E., Lobanova, E., Dinan, E., Smith, E. M., Radenovic, F., Guzmán, F., Zhang, F., Synnaeve, G., Lee, G., Anderson, G. L., Thattai, G., Nail, G., Mialon, G., Pang, G., Cucurell, G., Nguyen, H., Korevaar, H., Xu, H., Touvron, H., Zarov, I., Ibarra, I. A., Kloumann, I., Misra, I., Evtimov, I., Zhang, J., Copet, J., Lee, J., Geffert, J., Vranes, J., Park, J., Mahadeokar, J., Shah, J., van der Linde, J., Billock, J., Hong, J., Lee, J., Fu, J., Chi, J., Huang, J., Liu, J., Wang, J., Yu, J., Bitton, J., Spisak, J., Park, J., Rocca, J., Johnstun, J., Saxe, J., Jia, J., Alwala, K. V., Prasad, K., Upasani, K., Plawiak, K., Li, K., Heafield, K., Stone, K., El-Arini, K., Iyer, K., Malik, K., Chiu, K., Bhalla, K., Lakhotia, K., Rantala-Yeary, L., van der Maaten, L., Chen, L., Tan, L., Jenkins, L., Martin, L., Madaan, L., Malo, L., Blecher, L., Landzaat, L., de Oliveira, L., Muzzi, M., Pasupuleti, M., Singh, M., Paluri, M., Kardas, M., Tsimpoukelli, M., Oldham, M., Rita, M., Pavlova, M., Kambadur, M., Lewis, M., Si, M., Singh, M. K., Hassan, M., Goyal, N., Torabi, N.,

Bashlykov, N., Bogoychev, N., Chatterji, N., Zhang, N., Duchenne, O., Çelebi, O., Alrassy, P., Zhang, P., Li, P., Vasic, P., Weng, P., Bhargava, P., Dubal, P., Krishnan, P., Koura, P. S., Xu, P., He, Q., Dong, Q., Srinivasan, R., Ganapathy, R., Calderer, R., Cabral, R. S., Stojnic, R., Raileanu, R., Maheswari, R., Girdhar, R., Patel, R., Sauvestre, R., Polidoro, R., Sumbaly, R., Taylor, R., Silva, R., Hou, R., Wang, R., Hosseini, S., Chennabasappa, S., Singh, S., Bell, S., Kim, S. S., Edunov, S., Nie, S., Narang, S., Raparthy, S., Shen, S., Wan, S., Bhosale, S., Zhang, S., Vandenhende, S., Batra, S., Whitman, S., Sootla, S., Collot, S., Gururangan, S., Borodinsky, S., Herman, T., Fowler, T., Sheasha, T., Georgiou, T., Scialom, T., Speckbacher, T., Mihaylov, T., Xiao, T., Karn, U., Goswami, V., Gupta, V., Ramanathan, V., Kerkez, V., Gonguet, V., Do, V., Vogeti, V., Albiero, V., Petrovic, V., Chu, W., Xiong, W., Fu, W., Meers, W., Martinet, X., Wang, X., Wang, X., Tan, X. E., Xia, X., Xie, X., Jia, X., Wang, X., Goldschlag, Y., Gaur, Y., Babaei, Y., Wen, Y., Song, Y., Zhang, Y., Li, Y., Mao, Y., Coudert, Z. D., Yan, Z., Chen, Z., Papakipos, Z., Singh, A., Srivastava, A., Jain, A., Kelsey, A., Shajnfeld, A., Gangidi, A., Victoria, A., Goldstand, A., Menon, A., Sharma, A., Boesenberg, A., Baevski, A., Feinstein, A., Kallet, A., Sangani, A., Teo, A., Yunus, A., Lupu, A., Alvarado, A., Caples, A., Gu, A., Ho, A., Poulton, A., Ryan, A., Ramchandani, A., Dong, A., Franco, A., Goyal, A., Saraf, A., Chowdhury, A., Gabriel, A., Bharambe, A., Eisenman, A., Yazdan, A., James, B., Maurer, B., Leonhardi, B., Huang, B., Loyd, B., Paola, B. D., Paranjape, B., Liu, B., Wu, B., Ni, B., Hancock, B., Wasti, B., Spence, B., Stojkovic, B., Gamido, B., Montalvo, B., Parker, C., Burton, C., Mejia, C., Liu, C., Wang, C., Kim, C., Zhou, C., Hu, C., Chu, C.-H., Cai, C., Tindal, C., Feichtenhofer, C., Gao, C., Civin, D., Beaty, D., Kreymer, D., Li, D., Adkins, D., Xu, D., Testuggine, D., David, D., Parikh, D., Liskovich, D., Foss, D., Wang, D., Le, D., Holland, D., Dowling, E., Jamil, E., Montgomery, E., Presani, E., Hahn, E., Wood, E., Le, E.-T., Brinkman, E., Arcaute, E., Dunbar, E., Smothers, E., Sun, F., Kreuk, F., Tian, F., Kokkinos, F., Ozgenel, F., Caggioni, F., Kanayet, F., Seide, F., Florez, G. M., Schwarz, G., Badeer, G., Swee, G., Halpern, G., Herman, G., Sizov, G., Guangyi, Zhang, Lakshminarayanan, G., Inan, H., Shojanazeri, H., Zou, H., Wang, H., Zha, H., Habeeb, H., Rudolph, H., Suk, H., Aspegren, H., Goldman, H., Zhan, H., Damlaj, I., Molybog, I., Tufanov, I., Leontiadis, I., Veliche, I.-E., Gat, I., Weissman, J., Geboski, J., Kohli, J., Lam, J., Asher, J., Gaya, J.-B., Marcus, J., Tang, J., Chan, J., Zhen, J., Reizenstein, J., Teboul, J., Zhong, J., Jin, J., Yang, J., Cummings, J., Carvill, J., Shepard, J., McPhie, J., Torres, J., Ginsburg, J., Wang, J., Wu, K., U, K. H., Saxena, K., Khandelwal, K., Zand, K., Matosich, K., Veeraraghavan, K., Michelena, K., Li, K., Jagadeesh, K., Huang, K., Chawla, K., Huang, K., Chen, L., Garg, L., A, L., Silva, L., Bell, L., Zhang, L., Guo, L., Yu, L., Moshkovich, L., Wehrstedt, L., Khabsa, M., Avalani, M., Bhatt, M., Mankus, M., Hasson, M., Lennie, M., Reso, M., Groshev, M., Naumov, M., Lathi, M., Keneally, M., Liu, M., Seltzer, M. L., Valko, M., Restrepo, M., Patel, M., Vyatskov, M., Samvelyan, M., Clark, M., Macey, M., Wang, M., Hermoso, M. J., Metanat, M., Rastegari, M., Bansal, M., Santhanam, N., Parks, N., White, N., Bawa, N., Singhal, N., Egebo, N., Usunier, N., Mehta, N., Laptev, N. P., Dong, N., Cheng, N., Chernoguz, O., Hart, O., Salpekar, O., Kalinli, O., Kent, P., Parekh, P., Saab, P., Balaji, P., Rittner, P., Bontrager, P., Roux, P., Dollar, P., Zvyagina, P., Ratanchandani, P., Yuvraj, P., Liang, Q., Alao, R., Rodriguez, R., Ayub, R., Murthy, R., Nayani, R., Mitra, R., Parthasarathy, R., Li, R., Hogan, R., Battey, R., Wang, R., Howes, R., Rinott, R., Mehta, S., Siby, S., Bondu, S. J., Datta, S., Chugh, S., Hunt, S., Dhillon, S., Sidorov, S., Pan, S., Mahajan, S., Verma, S., Yamamoto, S., Ramaswamy, S., Lindsay, S., Lindsay, S., Feng, S., Lin, S., Zha, S. C., Patil, S., Shankar, S., Zhang, S., Zhang, S., Wang, S., Agarwal, S., Sajuyigbe, S., Chintala, S., Max, S., Chen, S., Kehoe, S., Satterfield, S., Govindaprasad, S., Gupta, S., Deng, S., Cho, S., Virk, S., Subramanian, S., Choudhury, S., Goldman, S., Remez, T., Glaser, T., Best, T., Koehler, T., Robinson, T., Li, T., Zhang, T., Matthews, T., Chou, T., Shaked, T., Vontimitta, V., Ajayi, V., Montanez, V., Mohan, V., Kumar, V. S., Mangla, V., Ionescu, V., Poenaru, V., Mihailescu, V. T., Ivanov, V., Li, W., Wang, W., Jiang, W., Bouaziz, W., Constable, W., Tang, X., Wu, X., Wang, X., Wu, X., Gao, X., Kleinman, Y., Chen, Y., Hu, Y., Jia, Y., Qi, Y., Li, Y., Zhang, Y., Zhang, Y., Adi, Y., Nam, Y., Yu, Wang, Zhao, Y., Hao, Y., Qian, Y., Li, Y., He, Y., Rait, Z., DeVito, Z., Rosnbrick, Z., Wen, Z., Yang, Z., Zhao, Z., and Ma, Z. The Llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783.

Guan, M. Y., Wang, M., Carroll, M., Dou, Z., Wei, A. Y., Williams, M., Arnav, B., Huizinga, J., Kivlichan, I., Glaese, M., Pachocki, J., and Baker, B. Monitoring monitorability, 2025. URL https://arxiv.org/abs/2512.18311.

Hase, P. and Potts, C. Counterfactual simulation training for chain-of-thought faithfulness, 2026. URL https://arxiv.org/abs/2602.20710.

Huang, V., Choi, D., Johnson, D. D., Schwettmann, S., and Steinhardt, J. Predictive concept decoders: Training scalable end-to-end interpretability assistants, 2025. URL https://arxiv.org/abs/2512.15712.

Ibrahim, L., Hafner, F. S., and Rocher, L. Training language models to be warm can reduce accuracy and increase sycophancy. *Nature*, 652(8112):1159–1165, Apr 2026. ISSN 1476-4687. doi: 10.1038/

s41586-026-10410-0. URL https://doi.org/10.1038/s41586-026-10410-0.

Jiang, L., Rao, K., Han, S., Ettinger, A., Brahman, F., Kumar, S., Mireshghallah, N., Lu, X., Sap, M., Choi, Y., and Dziri, N. Wildteaming at scale: From in-the-wild jailbreaks to (adversarially) safer language models. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=n5R6TvBVcX.

Joglekar, M., Chen, J., Wu, G., Yosinski, J., Wang, J., Barak, B., and Glaese, A. Training llms for honesty via confessions, 2025. URL https://arxiv.org/abs/2512.08093.

Karvonen, A., Chua, J., Dumas, C., Fraser-Taliente, K., Kantamneni, S., Minder, J., Ong, E., Sharma, A. S., Wen, D., Evans, O., and Marks, S. Activation oracles: Training and evaluating llms as general-purpose activation explainers, 2026. URL https://arxiv.org/abs/2512.15674.

Korbak, T., Balesni, M., Barnes, E., Bengio, Y., Benton, J., Bloom, J., Chen, M., Cooney, A., Dafoe, A., Dragan, A., Emmons, S., Evans, O., Farhi, D., Greenblatt, R., Hendrycks, D., Hobbhahn, M., Hubinger, E., Irving, G., Jenner, E., Kokotajlo, D., Krakovna, V., Legg, S., Lindner, D., Luan, D., Mądry, A., Michael, J., Nanda, N., Orr, D., Pachocki, J., Perez, E., Phuong, M., Roger, F., Saxe, J., Shlegeris, B., Soto, M., Steinberger, E., Wang, J., Zaremba, W., Baker, B., Shah, R., and Mikulik, V. Chain of thought monitorability: A new and fragile opportunity for AI safety, 2025. URL https://arxiv.org/abs/2507.11473.

Laine, R., Chughtai, B., Betley, J., Hariharan, K., Scheurer, J., Balesni, M., Hobbhahn, M., Meinke, A., and Evans, O. Me, myself, and ai: The situational awareness dataset (sad) for llms, 2024. URL https://arxiv.org/abs/2407.04694.

Lanham, T., Chen, A., Radhakrishnan, A., Steiner, B., Denison, C., Hernandez, D., Li, D., Durmus, E., Hubinger, E., Kernion, J., Lukošiūtė, K., Nguyen, K., Cheng, N., Joseph, N., Schiefer, N., Rausch, O., Larson, R., McCandlish, S., Kundu, S., Kadavath, S., Yang, S., Henighan, T., Maxwell, T., Telleen-Lawton, T., Hume, T., Hatfield-Dodds, Z., Kaplan, J., Brauner, J., Bowman, S. R., and Perez, E. Measuring faithfulness in chain-of-thought reasoning, 2023. URL https://arxiv.org/abs/2307.13702.

Lederman, H. and Mahowald, K. Emergent introspection in ai is content-agnostic, 2026. URL https://arxiv.org/abs/2603.05414.

Li, B. Z., Guo, Z. C., Huang, V., Steinhardt, J., and Andreas, J. Training language models to explain their own computations, 2025a. URL https://arxiv.org/abs/2511.08579.

Li, C., Phuong, M., and Tan, D. Spilling the beans: Teaching llms to self-report their hidden objectives, 2026. URL https://arxiv.org/abs/2511.06626.

Li, M., Arroyo, A. M. C., Rogers, G., Saphra, N., and Wallace, B. C. Do natural language descriptions of model activations convey privileged information? In *Mechanistic Interpretability Workshop at NeurIPS 2025*, 2025b. URL https://openreview.net/forum?id=zyhibAkzSA.

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

Macar, U., Yang, L., Wang, A., Wallich, P., Ameisen, E., and Lindsey, J. Mechanisms of introspective awareness, 2026. URL https://arxiv.org/abs/2603.21396.

Merullo, J., Eickhoff, C., and Pavlick, E. Circuit component reuse across tasks in transformer language models, 2024. URL https://arxiv.org/abs/2310.08744.

Pan, A., Chen, L., and Steinhardt, J. LatentQA: Teaching LLMs to decode activations into natural language. *arXiv*, 2024.

Pearson-Vogel, T., Vanek, M., Douglas, R., and Kulveit, J. Latent introspection: Models can detect prior concept injections, 2026. URL https://arxiv.org/abs/2602.20031.

Penedo, G., Kydlíček, H., allal, L. B., Lozhkov, A., Mitchell, M., Raffel, C., Werra, L. V., and Wolf, T. The fineweb datasets: Decanting the web for the finest text data at scale. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2024. URL https://openreview.net/forum?id=n6SCkn2QaG.

Plunkett, D., Morris, A., Reddy, K., and Morales, J. Self-interpretability: Llms can describe complex internal processes that drive their decisions, 2025. URL https://arxiv.org/abs/2505.17120.

Pres, I., Li, B. Z., Ruis, L., Guo, Z. C., Hu, K., Damani, M., Puri, I., Lubana, E. S., and Andreas, J. Position: It's time to optimize for self-consistency. In *Forty-third International Conference on Machine Learning Position Paper Track*, 2026. URL https://time-for-consistency.github.io/.

Sharma, M., Tong, M., Korbak, T., Duvenaud, D., Askell, A., Bowman, S. R., Cheng, N., Durmus, E., Hatfield-Dodds, Z., Johnston, S. R., Kravec, S., Maxwell, T., McCandlish, S., Ndousse, K., Rausch, O., Schiefer, N., Yan, D., Zhang, M., and Perez, E. Towards understanding sycophancy in language models, 2025. URL https://arxiv.org/abs/2310.13548.

Shenoy, K., Yang, L., Sheshadri, A., Mindermann, S., Lindsey, J., Marks, S., and Wang, R. Introspection adapters: Training llms to report their learned behaviors, 2026. URL https://arxiv.org/abs/2604.16812.

Sheshadri, A., Ewart, A., Guo, P. H., Lynch, A., Wu, C., Hebbar, V., Sleight, H., Stickland, A. C., Perez, E., Hadfield-Menell, D., and Casper, S. Latent adversarial training improves robustness to persistent harmful behaviors in LLMs, 2025. URL https://openreview.net/forum?id=wI5uHZLeCZ.

Song, S., Hu, J., and Mahowald, K. Language models fail to introspect about their knowledge of language. In *Second Conference on Language Modeling*, 2025a. URL https://openreview.net/forum?id=AivRDOFi5H.

Song, S., Lederman, H., Hu, J., and Mahowald, K. Privileged self-access matters for introspection in AI, 2025b. URL https://arxiv.org/abs/2508.14802.

Turpin, M., Michael, J., Perez, E., and Bowman, S. R. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=bzs4uPLXvi.

Wang, K. R., Variengien, A., Conmy, A., Shlegeris, B., and Steinhardt, J. Interpretability in the wild: a circuit for indirect object identification in GPT-2 small. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=NpsVSN6o4ul.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., brian ichter, Xia, F., Chi, E. H., Le, Q. V., and Zhou, D. Chain of thought prompting elicits reasoning in large language models. In Oh, A. H., Agarwal, A., Belgrave, D., and Cho, K. (eds.), *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=_VjQlMeSB_J.

Wei, J., Huang, D., Lu, Y., Zhou, D., and Le, Q. V. Simple synthetic data reduces sycophancy in large language models, 2024. URL https://arxiv.org/abs/2308.03958.

Yang, A., Li, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Gao, C., Huang, C., Lv, C., Zheng, C., Liu, D., Zhou, F., Huang, F., Hu, F., Ge, H., Wei, H., Lin, H., Tang, J., Yang, J., Tu, J., Zhang, J., Yang, J., Yang, J., Zhou, J., Zhou, J., Lin, J., Dang, K., Bao, K., Yang, K., Yu, L., Deng, L., Li, M., Xue, M., Li, M., Zhang, P., Wang, P., Zhu, Q., Men, R., Gao, R., Liu, S., Luo, S., Li, T., Tang, T., Yin, W., Ren, X., Wang, X., Zhang, X., Ren, X., Fan, Y., Su, Y., Zhang, Y., Zhang, Y., Wan, Y., Liu, Y., Wang, Z., Cui, Z., Zhang, Z., Zhou, Z., and Qiu, Z. Qwen3 technical report, 2025. URL https://arxiv.org/abs/2505.09388.

Zhang, Z., Xu, W., Wu, F., and Reddy, C. K. Falsereject: A resource for improving contextual safety and mitigating over-refusals in LLMs via structured reasoning. In *Second Conference on Language Modeling*, 2025. URL https://openreview.net/forum?id=1w9Hay7tvm.

Zhao, W., Ren, X., Hessel, J., Cardie, C., Choi, Y., and Deng, Y. Wildchat: 1m chatgpt interaction logs in the wild. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. URL https://openreview.net/forum?id=Bl8u7ZRlbM.

Zhong, Z. Spontaneous introspection in output tampering. LessWrong, April 2026. URL https://www.lesswrong.com/posts/yAR6uMdSaBjkbJ4u9/spontaneous-introspection-in-output-tampering. Accessed: 2026-05-06.

A. Do model behavior shifts to become easier to explain?

A.1. Fine-grained metrics

In the main paper, we evaluate the model's explanation on one main metric, the Explanation Exact Match (EM). We provide more fine-grained metrics to supplement here. For each example $\hat{E}(M)(x, C)$ we check two conditions:

- *flag match*: $\texttt{is_changed}(\hat{E}(M)(x, C)) = \mathbf{1}[\, B(M)(x) \neq B(M)(x_{\setminus C})\,]$ — the predicted change-flag agrees with whether M's behavior actually changes under cue ablation. From this, we can evaluate the F1 score for the two categories, "changed" and "unchanged."
- *content match*: $\hat{y}_{\setminus C}(\hat{E}(M)(x, C)) = B(M)(x_{\setminus C})$ — the predicted post-ablation output equals M's actual cue-ablated behavior.

The explanation evaluation for all the metrics can be seen in Figure 9. Behavior panel 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 $\mathcal{M}_0$. [5] Shown in the Explanation panel, 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.

A.2. Training on external model on $\mathcal{M}_{\text{reg}}$'s behavior distribution

Is $E(\mathcal{M}_{\text{reg}})$ just an easier distribution? 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 performing self-introspection. We verify that this is not the case by training another model, derived from base Qwen3-8B rather than $\mathcal{M}_{\text{reg}}$, on $E(\mathcal{M}_{\text{reg}})$, that we call $\mathcal{M}^{(2)}_{\text{reg}} = M[B(\mathcal{M}_{\text{reg}}), E(\mathcal{M}_{\text{reg}})]$. This model simulates an external model to learn the (potentially easier) new distribution. We find two pieces of evidence against this: (1) In Figure 9, $\mathcal{M}^{(2)}_{\text{reg}}$ explaining $\mathcal{M}_{\text{reg}}$ (purple bar) is not better than $\mathcal{M}_{\text{reg}}$ explaining $\mathcal{M}_0$ (blue bar), meaning that $E(\mathcal{M}_{\text{reg}})$ is not an easier distribution to learn. (2) $\mathcal{M}^{(2)}_{\text{reg}}$ explaining $\mathcal{M}_{\text{reg}}$ (purple bar) is not better than $\mathcal{M}_{\text{reg}}$ explaining $\mathcal{M}_{\text{reg}}$ (green bar), meaning that self explainer is better than an external explainer. More surprisingly, for $\mathcal{M}^{(2)}_{\text{reg}}$, **Self > Orig** persists, i.e. $\hat{E}(\mathcal{M}^{(2)}_{\text{reg}})$ matches $E(\mathcal{M}^{(2)}_{\text{reg}})$ better than $E(\mathcal{M}_{\text{reg}})$ for this new model (Figure 9). The gap is not an artifact of the original explainer being closer to itself by construction.

[5] 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.

B. Training dynamics

We track when introspective coupling emerges during training and conversely, when self- and original-explanation decouple. At every training batch, we record three quantities: (i) on behavior data, whether $B(\mathcal{M})$ matches the training label $B(\mathcal{M}_0)$; (ii) on the explanation task, whether $\hat{E}(\mathcal{M})$ matches the training label $E(\mathcal{M}_0)$; and (iii) on the explanation task, whether $\hat{E}(\mathcal{M})$ matches the explanation $E(\mathcal{M})$ constructed from the model's own current behavior.

We smooth each per-batch series with a time-weighted exponential moving average[6] and report the resulting curves for $\mathcal{M}_{\text{reg}}$ and $\mathcal{M}_{\text{unreg}}$ in Figure 10(a) and (b) respectively, with panel (c) plotting the per-step self-minus-orig explanation delta for both models. With regularization, $\mathcal{M}_{\text{reg}}$'s self- and orig-explanation curves remain tightly coupled throughout training, with self-explanation gradually pulling ahead. Without regularization, $B(\mathcal{M})$ drifts sharply from $B(\mathcal{M}_0)$ within the first 1000 steps—almost entirely on the cue-included answer. With this behavioral drift, the self-vs-orig explanation gap widens accordingly.

C. Only Higher-Rank LoRA Recovers Self > Orig

We default to using full-finetuning (FFT) in the main paper as it more consistently presents the Self > Orig gap for investigative purposes. Here, we show that, specifically for the HINT-MMLU task, only sufficiently high-rank LoRA adapters can recover the self > orig effect. We sweep along the matched-scaling diagonal $\alpha = r$ for $r \in \{32, 64, 80, 96, 128, 256\}$ and report the Behavior EM and Explanation EM in Figure 11. In the figure, we show that for all LoRA adapters, the behavioral EM always drifts, but only sufficiently high-rank LoRA adapters exhibit Self > Orig on explanation. Therefore, the low-rank LoRA adapters' inability to elicit the phenomenon cannot be attributed to the fact that they don't drift as much, but rather that they genuinely cannot distinguish the two labels.

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

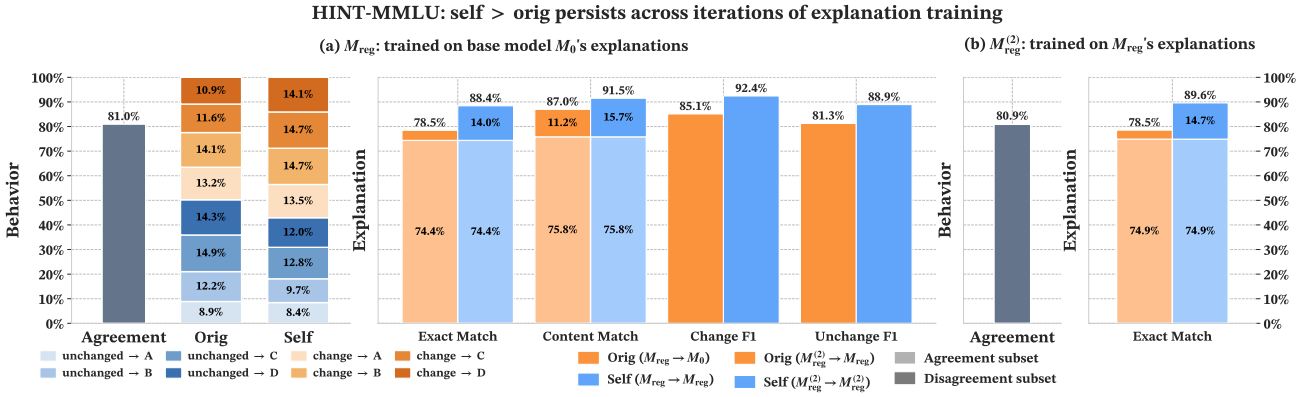


Figure 9. **(a)** Regularized explainer $\mathcal{M}_{\text{reg}}$ trained on $\mathcal{M}_0$'s explanations. Behavior breakdown shows the distribution of 8-label breakdown for $B(\mathcal{M}_0)$, and that it doesn't drift into a degenerate case. Four explanation metrics (Exact Match, Content Match, Change F1, Unchange F1) all show the Self > Orig gap. **(b)** External explainer baseline $\mathcal{M}_{\text{reg}}^{(2)}$: 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 $\mathcal{M}_{\text{reg}}^{(2)}$.

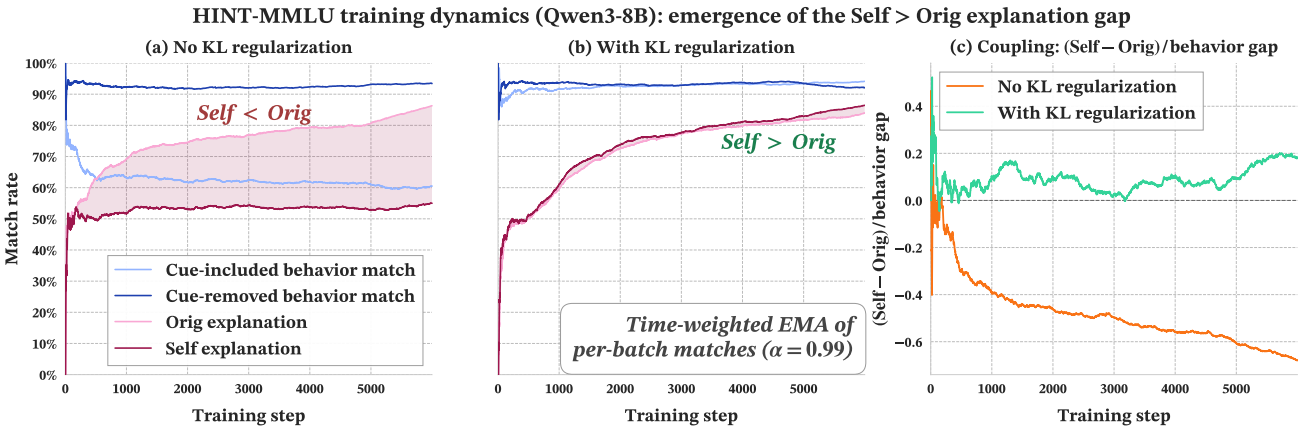


Figure 10. Training dynamics of Qwen3-8B on HINT-MMLU. **(a)** No regularization on behavior vs. **(b)** with KL regularization. We plot time-weighted exponential moving average with $\alpha = 0.99$ of the corresponding per-batch metric. Blue lines plot the two behavioral drift metrics, and the two explanation lines match explanation against the model's own behavior labels (purple) and against the original $\mathcal{M}_0$ labels (pink). **(c)** the per-step ratio of explanation (Self − 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.

D. Higher Learning Rate Can Widen The Gap

We default to using $lr = 1 \cdot 10^{(-5)}$ throughout the paper for training, where we find this to give the best absolute numbers on Explanation EM. However, here we perform a sweep on different learning rates and measure how the learning rate affects the emergence of the Self > Orig gap. We sweep six different learning rates ranging from $1 \cdot 10^{-4}$ to $5 \cdot 10^{-6}$ and plot the Behavioral and Explanation EM in Figure 12. We find that, with an increasingly bigger learning rate, the absolute explanation EM number drops on both Orig labels and Self labels, but it drops more quickly for Orig than Self and widens the Self > Orig gap. Behavior EM and Orig-side Explanation EM decay in lockstep.

Counter-evidence to the threshold hypothesis. The LR sweep is the one experiment in this paper that the simple $E \leftrightarrow M$ divergence threshold does *not* explain. At very large learning rates the trained model drifts off base by more than 70% on Behavior EM — well past the $\tau \approx 0.3$ threshold derived in Section 4 — and yet Self > Orig remains present and even grows. Symmetrically, at $\text{lr} = 5\text{e}{-6}$ the model barely drifts at all, which on the threshold story should be the easiest regime in which to recover Self > Orig, but the gap is in fact *absent* there. We flag this inversion as an open puzzle: a plausible refinement is that the LR also controls how aggressively the explanation head can re-route through the model's own current circuitry (independent of behavioral drift magnitude), but we have not yet pinned this down empirically.

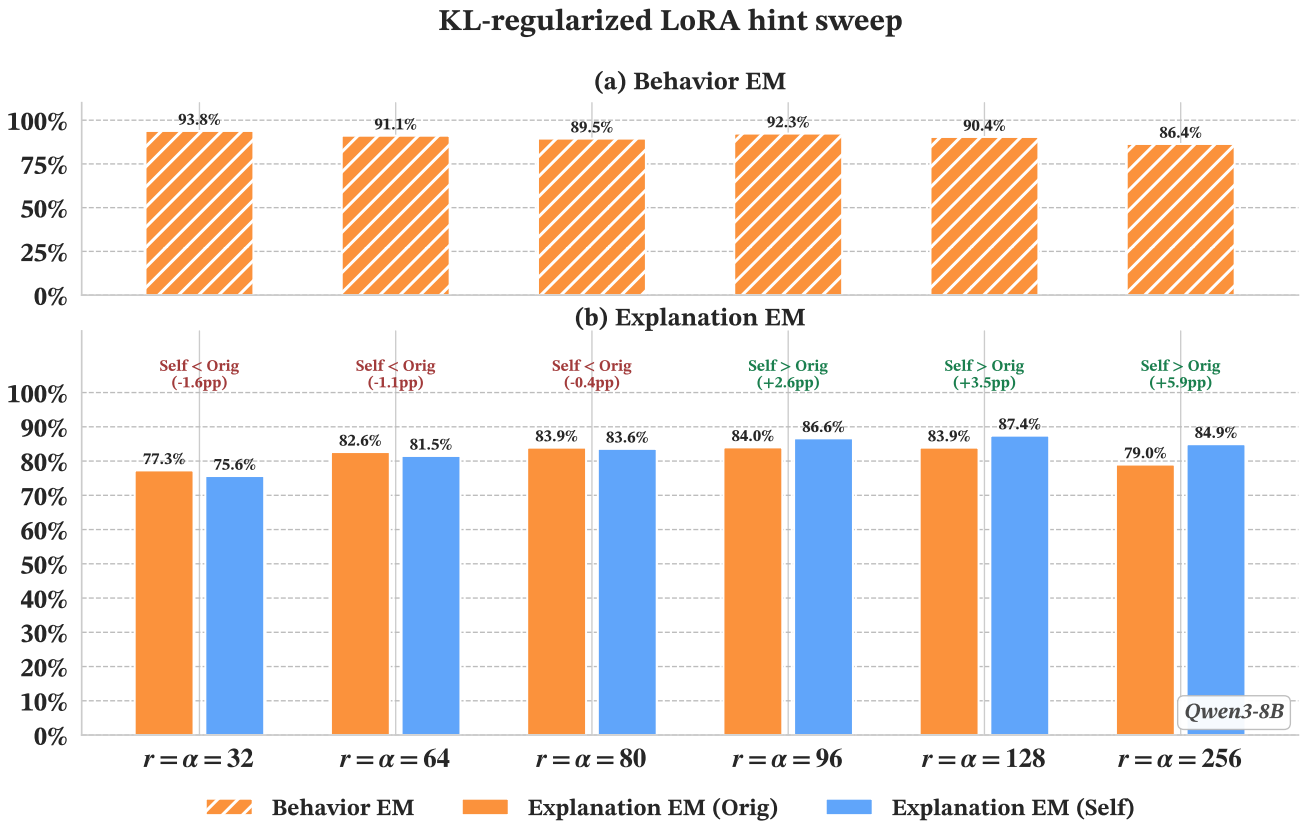


Figure 11. KL-regularized LoRA rank sweep on HINT-MMLU, $\alpha = r$. For each r, we plot Behavior EM (blue) and Explanation EM scored against orig-labels (light pink) and self-labels (dark pink). The self > orig gap is absent or slightly negative for $r \leq 80$, opens at $r = 96$, and widens monotonically through $r = 256$.

E. Detailed Explanation Results

E.1. Untrained baseline

Following (Li et al., 2025a), 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 13. Across all three tasks the untrained few-shot baseline reaches only 14–18% Explanation EM, far below the trained explainer, establishing that self-explanation is not a native capability of $\mathcal{M}_0$.

E.2. AITA Results

For the AITA setup (Cheng et al., 2026), training data is derived from this dataset `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.
```

The self-vs-orig metrics decomposition for the model trained on the AITA dataset is shown in Figure 14. Note that the test set is balanced over the unhinted-condition NTA/YTA verdict, rather than the change/unchange verdict. We show that self-explanation is better on all four metrics for explanation, and especially recovers the gap between Change F1 and Unchange F1 on the Orig set.

E.3. Refusal Results

We adopt the following style of roleplay system prompt to nudge the model towards refusal or compliance, generated by gpt-5. Here are two examples:

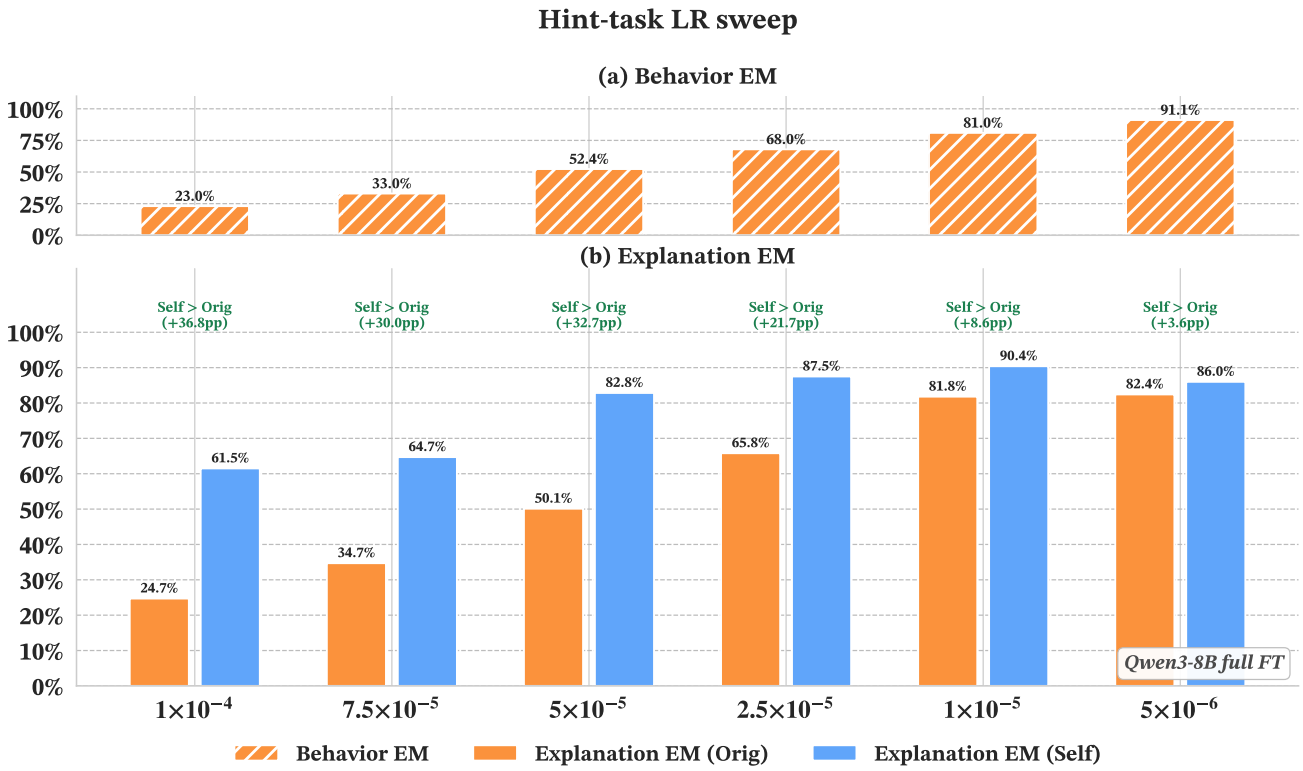


Figure 12. LR sweep on HINT-MMLU plotted with learning rate decreasing left-to-right. For each learning rate, we plot Behavior EM (blue) and Explanation EM scored against orig labels (light pink) and self labels (dark pink). Self > Orig is absent at $\mathrm{lr} = 5\mathrm{e}{-6}$ and continues to widen as the learning rate rises and the model drifts further from base.

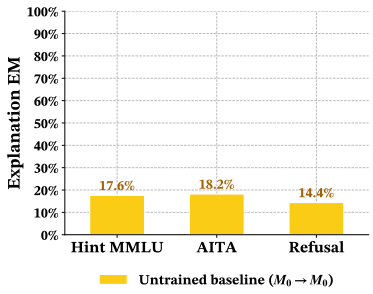


Figure 13. Untrained few-shot baseline ($\mathcal{M}_0 \to \mathcal{M}_0$) Explanation EM on each of the three tasks. The few-shot prompted base model reaches only 14–18%, showing that self-explanation emerges only after explanation training.

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.
```

We adopt the model's natural behavioral distribution. Because of alignment post-training, the label of "flipping to refusal" under any system prompt is extremely rare in its training data, so explanation training cannot learn that label. This is one of our limitations, that we need high-variance counterfactuals in the training data for the explanation not to be trivial.

The self-vs-orig metrics decomposition for the refusal explanation training is shown in Figure 15. We show that self explanation is much better than explaining the original model's behaviors on all aspects.

E.4. 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 HINT-MMLU run with behavioral regularization on Llama-3.1-8B-Instruct (Grattafiori et al., 2024) and Qwen3-32B (Yang et al., 2025). Because full fine-tuning is pro-

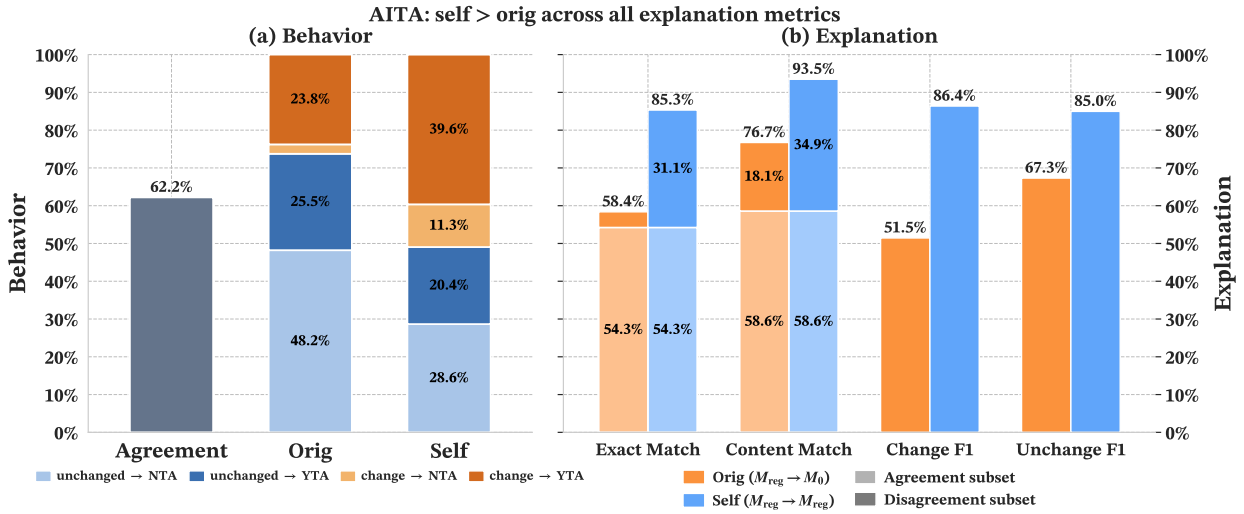


Figure 14. AITA $\mathcal{M}_{\text{reg}}$ detailed metrics. Behavior change rate (a) and four explanation-quality metrics (b) for the explainer trained on the original target's labels (indigo) vs. on self labels (teal). Each explanation bar is decomposed into the agreement-subset matched baseline (hatched) and the disagreement-subset gain (solid).

hibitive at 32B, we use **LoRA** ($r = \alpha = 128$), rather than the full-FT recipe used for the 8B runs. The full metrics decomposition is shown in Figures 16 and 17; the explainer does better on self-labels than orig-labels on every explanation-quality metric.

F. Detail for Auxiliary Training

F.1. Jabberwocky

We show an example of Jabberwocky data generated by an LLM following the HINT-MMLU format, but on fictitious names:

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
```

The main-text Jabberwocky figure (Figure 7) collapses behavior accuracy into two metrics, Behavior EM and Explanation EM. Figure 18 shows the full metrics: cue-included and cue-ablated match for behavior EM, and Exact Match, Content Match, Change F1, and Unchange F1 for explanation. On every metric, MMLU (Self) sits above MMLU (Orig), with high accuracy on the Jabberwocky set.

Figure 19 addresses a separate question on the Jabberwocky-v2 set itself ($n = 1732$): how does $\mathcal{M}_{\text{aux}}$'s introspection compare to a zero-shot $\mathcal{M}_0$ baseline, and is the gain a single-letter shortcut? The left panels show that $\mathcal{M}_0$ defaults to C on 63% of no-hint inputs while $\mathcal{M}_{\text{aux}}$ is approximately uniform, so the $0.06\% \rightarrow 79.85\%$ Explanation EM jump cannot be reproduced by always guessing a fixed letter. The right panel shows that the gain holds under letter-only scoring (semantic match: $23.8\% \rightarrow 84.2\%$) and that the residual gap concentrates in the unchanged class (Unchange F1 $0.000 \rightarrow 0.601$).

G. Detail for Behavioral Drift Auxiliary Training

Warm & Empathetic Assistant Response Following Ibrahim et al. (2026), we procure a corpus of ShareGPT-Vicuna user assistant conversations (Chiang et al., 2023) 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` (Sheshadri et al., 2025). Noticeably, direct refusal is a different type of behavior than that of typical Qwen3-8B, as it tends not to refuse directly, such as by saying "I'm sorry, but I cannot provide assistance with this request" but rather provides a long-winded justification. We mix this corpus of direct refusal training with training the model to counterfactually reason about whether the system prompt ablation would cause it to change its answer or not.

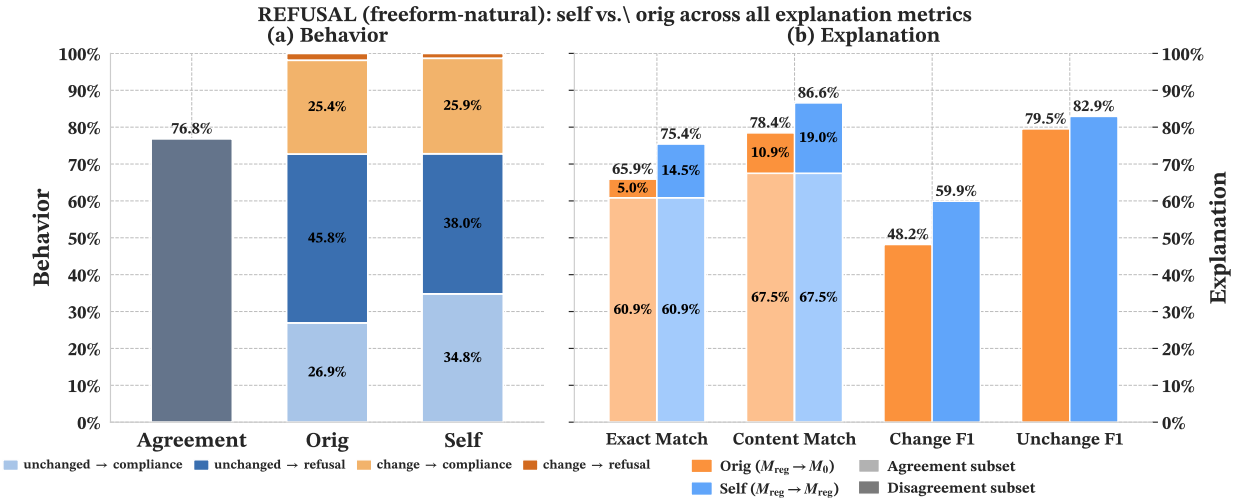


Figure 15. Refusal Explanation with Behavioral Regularization. Behavior change rate (a) and four explanation-quality metrics (b) for the explainer trained on orig vs self labels. Self > Orig on all metrics.

G.1. General Capabilities Evaluation

To check that explanation training mixing with general post-training pipelines does not erode general capabilities, we evaluate $\mathcal{M}_0$ and two trained variants on a general suite from `lm-evaluation-harness` (Gao et al., 2024): one model with $\mathcal{D}$ = Hint-MMLU, A = WildChat and another with $\mathcal{D}$ = Refusal Explanation, A = FineWeb. Table 2 reports absolute accuracy (%) and the delta against $\mathcal{M}_0$. Aggregate degradation is mild (within $\sim$2 pp on most tasks); the salient drops are GSM8K under both variants and TruthfulQA MC2 for the refusal explaining model.

H. Activation patching

H.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 causing changes in the explanation as well.

Figure 20a reports the same analysis on the model $\mathcal{M}_{\text{unreg}}$ trained without direct behavioral regularization. We show that the correlation is not as obvious: per-direction Pearson $r = 0.48/0.58$, combined $r = 0.53$ over the same activations of the hint region. Figure 20b shows that the base model $\mathcal{M}_0$ produces no correlation between the behavior and explanation at all.

H.2. Correlation between cue-ablated behavior and explanation

Section 3.3 reports the per-cell correlation between the explanation logit-diff and the *cue-included* answer-letter logit-diff. Here we run the same patching analysis with the object-level behavior metric replaced by the *cue-ablated* answer-letter logit-diff. 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 the hint region. Because the position to elicit the answer without hint comes before, we choose to patch in the last 10 tokens of the prompts. Note that this setup is noisier because the tokens we patch in are not shared between the two prompts of the pair.

Figure 21 shows that only patching a prompt with a "change" explanation into a prompt with an "unchange" explanation has a positive correlation, and patching a prompt with an "unchange" explanation into a prompt with a "change" explanation has no correlation at all. We compare this against the $\mathcal{M}_{\text{reg}}$ baseline, and show that there is no correlation in either direction. The reason behind why the correlation can only be observed on one side remains unanswered. The absolute normalized logit difference (NLD) for either direction is visibly smaller than that of the cue-included studies, with a maximum $NLD \approx 0.2$ vs. ≈ 0.4 for the cue-included behavior. This partially confirms that the setup is noisier.

I. Compute

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

J. Limitations

Cue ablation must produce non-saturated behavior flips. The training signal vanishes when $\mathcal{M}_0$'s behavior on x vs. $x_{\setminus C}$ is nearly always the same (or nearly always dif-

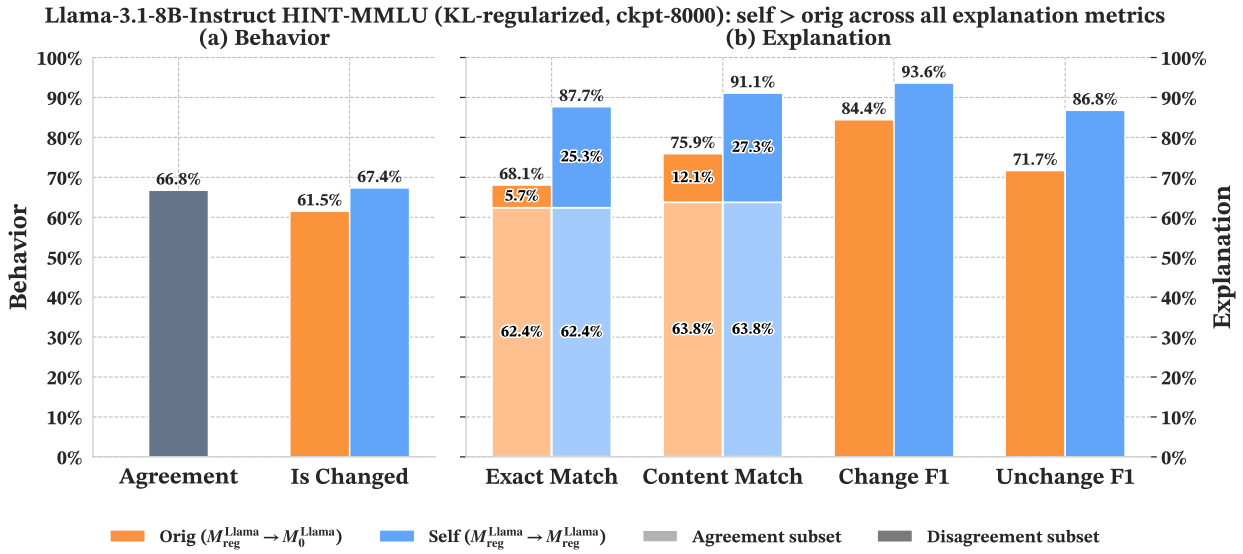


Figure 16. Llama-3.1-8B-Instruct trained and regularized on HINT-MMLU. Self > Orig on every explanation metric.

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)	`hint_wildchat` ($\mathcal{M}_{\text{reg}}$)	`fineweb_reg01`
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)

ferent): $E(()\mathcal{M}_0)$ collapses to a constant and explanation training degenerates into majority-class prediction. Heavily post-trained refusal is the canonical bad case — when alignment has driven $\mathcal{M}_0$ to refuse essentially every adversarial prompt, the change indicator is degenerate.

Behavior shifts must stay within a bounded band. Self > Orig requires $\mathcal{M}_{\text{reg}}$ to move toward the explanations it was trained on, but only modestly. With weak or no KL regularization, behavior drifts far enough from $\mathcal{M}_0$ that the trained explanation distribution no longer describes current behavior, and the self-explanation advantage disappears. Several auxiliary datasets in Section 5 (e.g. direct refusal dataset) materially shift $\mathcal{M}_0$'s behavior a lot more than when trained together with the explanation objective, as the KL regularization to $B(\mathcal{M}_0)$ damps this drift. We do not have an example of behavior shifting extremely drastically, and the explanation still tracks.

K. Broader Impact

Faithful introspection is a prerequisite for trustworthy 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. Conversely, our results show that naive supervised fine-tuning on explanations produces *decoupled* verbalizations that look introspective but are not, which risks giving operators a false sense of transparency and could be exploited to launder unsafe behavior behind plausible-sounding justifications. Verifying causal coupling could be beneficial before treating model-generated explanations as evidence about the model's true decision-making process.

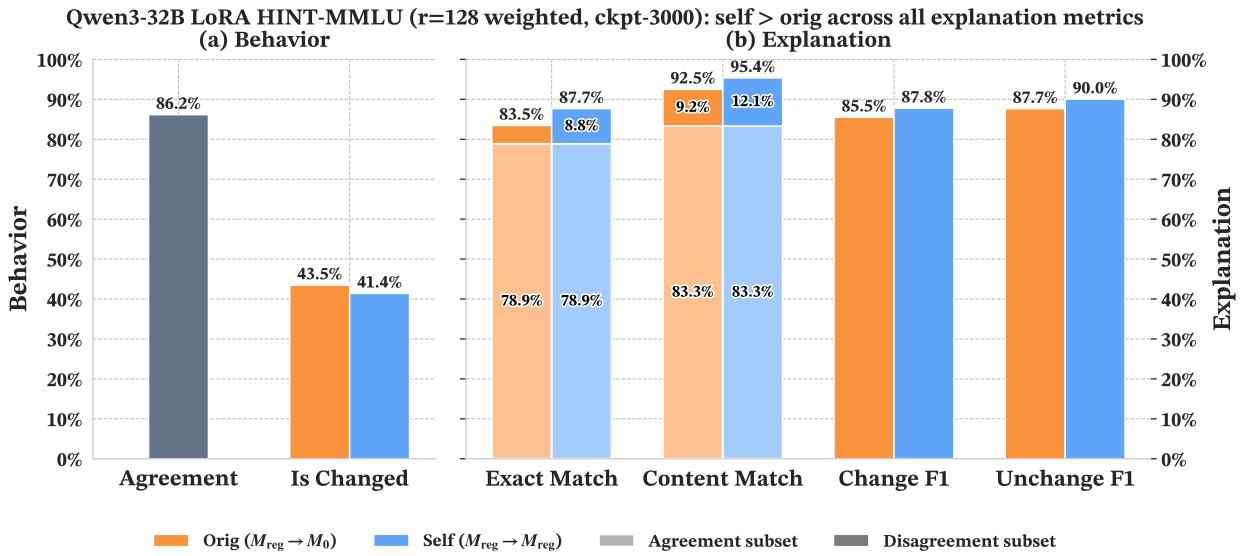


Figure 17. Qwen3-32B trained and regularized on HINT-MMLU (LoRA $r = \alpha = 128$). Self > Orig on every explanation metric.

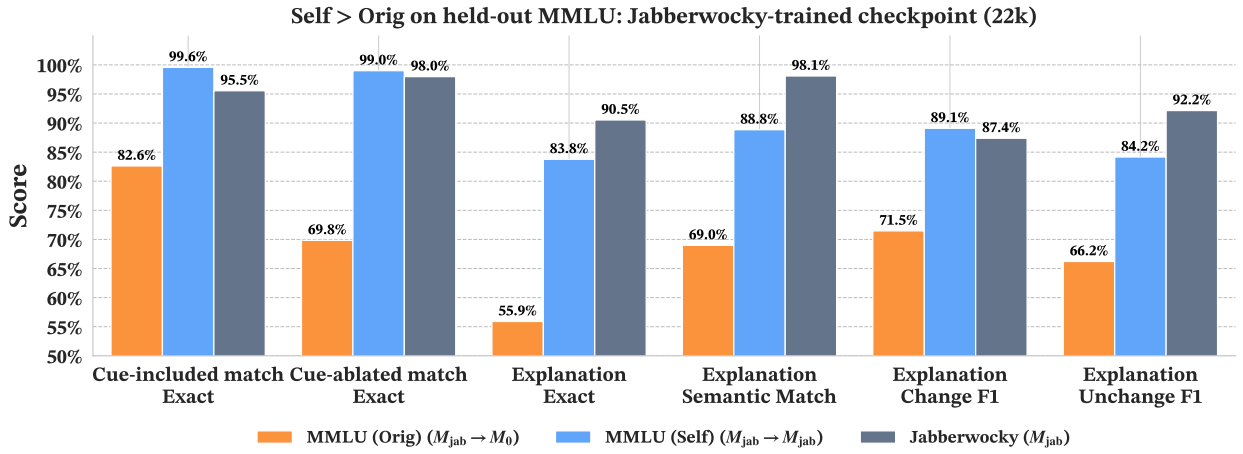


Figure 18. Full six-metric breakdown for Jabberwocky mixed training as Figure 7

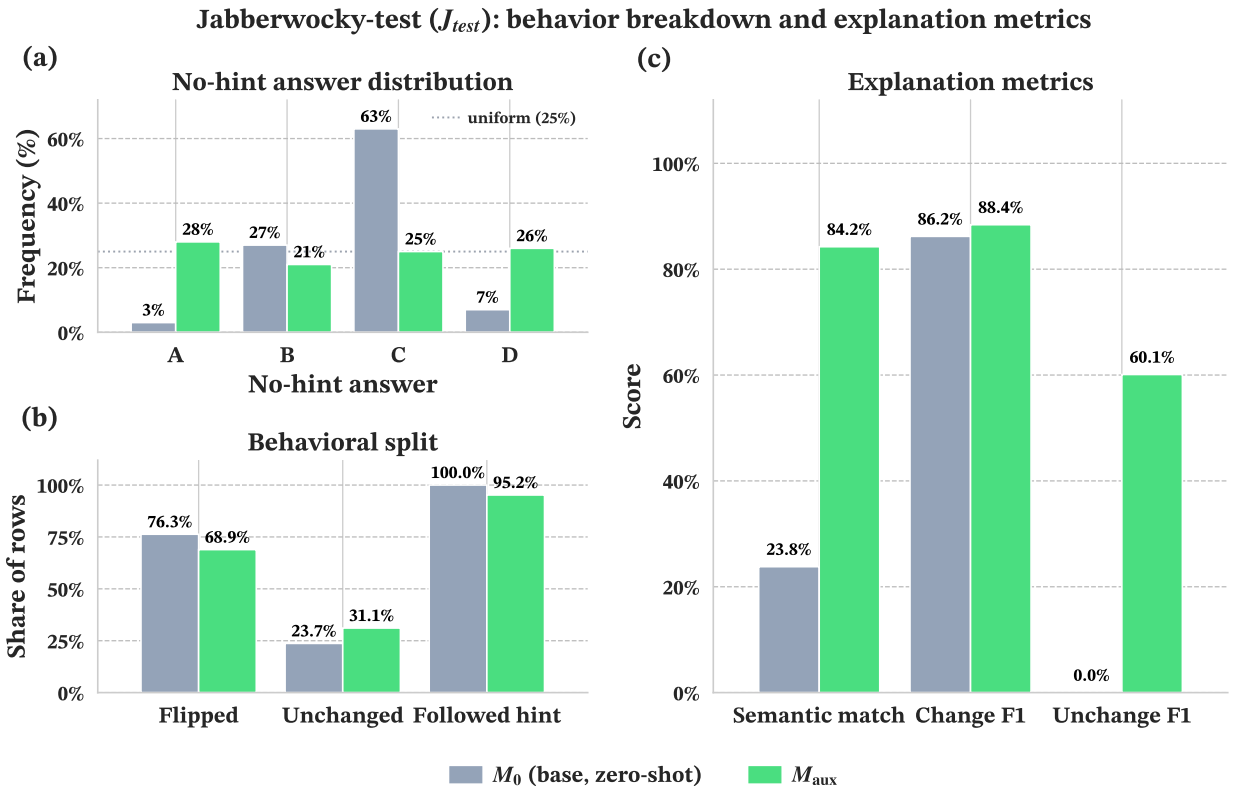
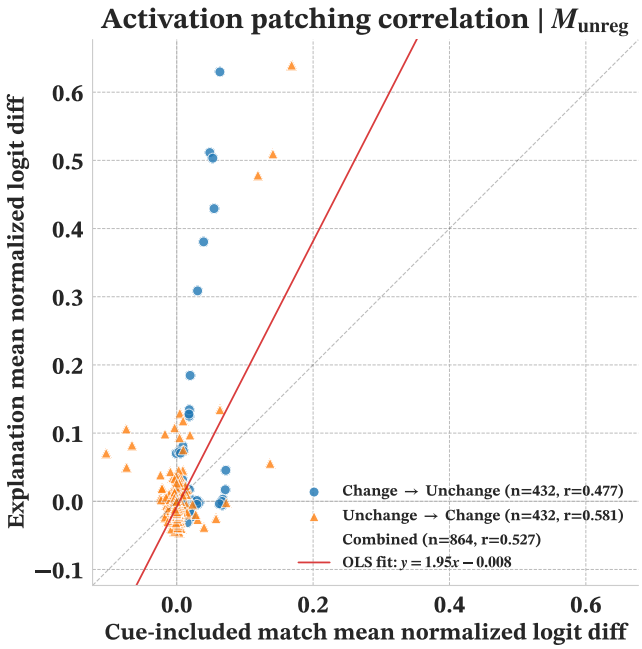
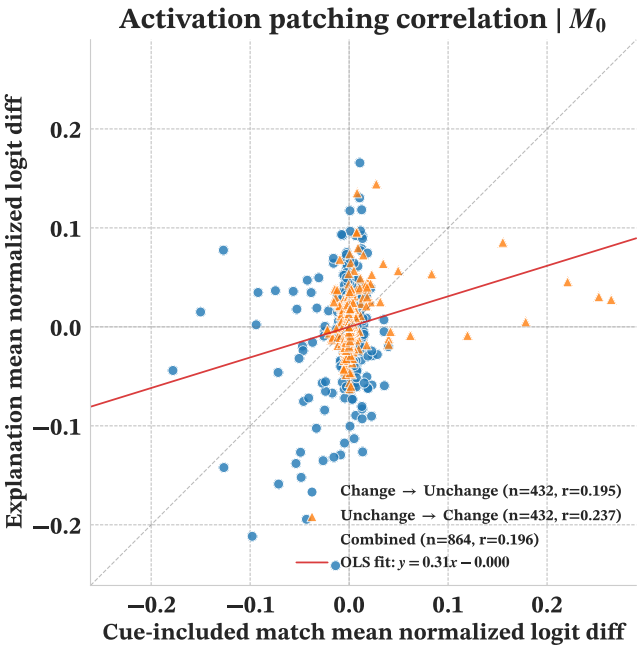


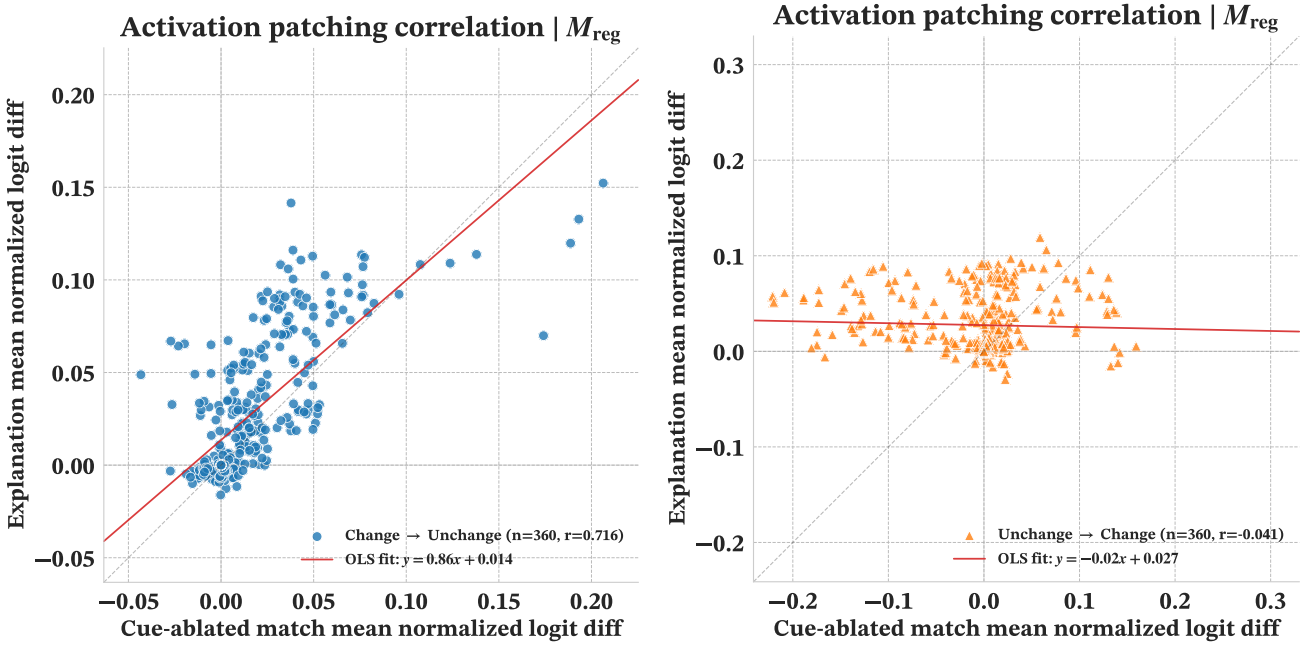
Figure 19. Jabberwocky-tset J_{test} detailed behavior and explanation metrics. **(a):** no-hint A/B/C/D answer distribution, showing that the trained $\mathcal{M}_{\text{aux}}$ is near random. **Bottom-left:** percentages for whether the answer is flipped, unchanged, or followed-hint. As the questions are nonsensical, the model defers to the hint. **Right:** explanation semantic letter 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.



(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 roughly halves the Pearson correlation between behavior and explanation logit-diffs.

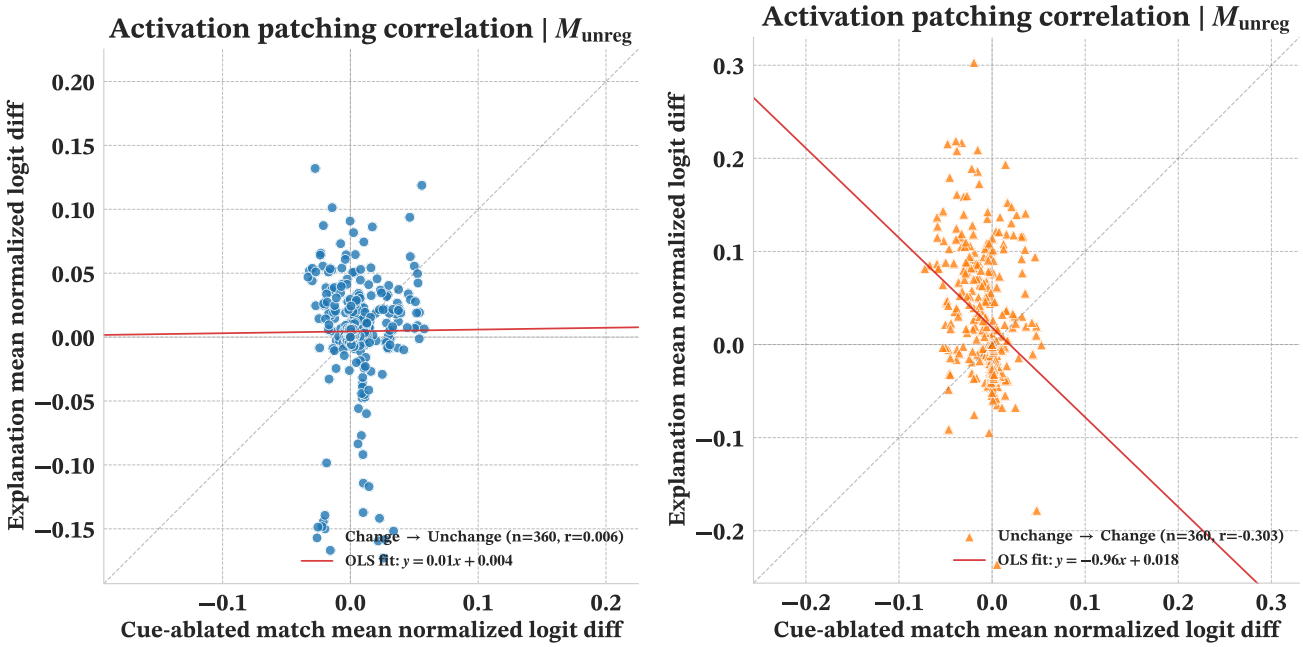


(b) $\mathcal{M}_0$ baseline counterpart to Figure 3: same per-(layer, token) scatter of mean normalized logit-diffs. $\mathcal{M}_0$ without explanation training has no correlation at all. Per-direction $r = 0.195/0.237$; combined $r = 0.196$.



(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 21. 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.