UNDERSTANDING DISTRESS IS NOT EXPRESSING EMPATHY: SUPPORTIVE EMPATHIC RESPONSE NEURONS IN LLMS

Anonymous authors
Paper under double-blind review

ABSTRACT

Large language models (LLMs) can recognize user distress and generate empathic responses. However, whether these behaviors rely on shared or separable internal mechanisms remains unclear. Because input-encoded information is not necessarily utilized during generation, we distinguish *distress representation* from *empathic response realization*. To isolate this response-side mechanism, we introduce a counterfactual difference-in-differences (DiD) activation framework. This framework contrasts matched supportive and neutral responses across distress and non-distress contexts. In this way, it mitigates confounds from input-side sensitivity and generic supportive language. The procedure identifies a sparse population of post-gating feed-forward network (FFN) units, termed *Supportive Empathic Response* (SER) neurons. Experiments demonstrate that SER neurons preferentially track affective empathic expression rather than instrumental support or generic supportive content, while remaining largely invariant to distress magnitude. Crucially, direct activation interventions across all four model families demonstrate reproducible bidirectional causal effects on both expressed empathy and support-strategy realization: suppressing frozen SER neurons reduces EPITOME-based expressed empathy by 10.3%–31.9% relative to baseline and shifts ESConv responses from affective toward instrumental support, whereas amplification consistently produces the opposite pattern. Furthermore, these interventions have only limited effects on general reasoning and broader emotion-related capabilities. Together, these findings identify a sparse, causally consequential response-side mechanism and provide evidence that representing user distress and realizing an empathic response are supported by partially dissociable internal mechanisms in LLMs.

1 INTRODUCTION

Large language models (LLMs) increasingly interact with users who disclose sadness, anxiety, loneliness, frustration, and other forms of distress. Responding appropriately in such settings requires not only recognizing the affective state of the user but also expressing that understanding in the response. A similar distinction appears in accounts of human empathy. Wiseman (1996) distinguishes understanding the feelings of another person from communicating that understanding, while appraisal theories separate the evaluation of an affectively relevant situation from subsequent emotional and coping responses (Lazarus, 1991). These accounts motivate a process-level distinction between representing another person's affective state and expressing that representation through an outward response.

For LLMs, this distinction raises a concrete mechanistic question. A model may recognize that a user is distressed yet respond with detached advice; conversely, it may produce fluent expressions of sympathy without adequately representing the situation that elicited the distress. We distinguish *distress representation* from *empathic response realization*: the former concerns the internal representation of the user's affective state, whereas the latter concerns how empathic understanding is expressed during response generation. Behavioral evidence of empathy alone does not establish that the same internal mechanism supports both processes. More generally, information recoverable from an internal representation need not be information that the model causally uses or expresses (Antverg

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

& Belinkov, 2022; Li et al., 2023). This leads to our central question: *can distress representation and empathic response realization be mechanistically separated within LLMs?*

Answering this question requires more than identifying neurons correlated with emotional dialogue. Two potential confounds must be distinguished. First, neurons that distinguish distressed from non-distressed contexts may primarily encode the user's affective state rather than contribute to response generation. Second, neurons that distinguish supportive from neutral responses may reflect generic reassurance, politeness, or supportive style rather than empathy conditioned on distress. A response-side mechanism of empathic realization should therefore be distinguishable from both *input-side distress sensitivity* and *generic supportive language*. Figure 1 illustrates this distinction: the user context remains fixed, while intervention on a response-side population changes how the model expresses support.

This motivates an identification strategy that holds the user context fixed while discounting response effects that also arise in non-distress contexts. We introduce a counterfactual difference-in-differences (DiD) framework that first contrasts matched supportive and neutral responses within the same context and then subtracts the corresponding contrast in non-distress contexts. The resulting statistic targets distress-conditioned response-side activity while subtracting effects shared with generic supportive language. The strongest statistically reliable post-gating feed-forward network (FFN) dimensions define a sparse population that we call *Supportive Empathic Response* (SER) neurons. The identified neuron sets are frozen before all held-out and downstream analyses and are compared with layer-matched control populations.

Identification is only the first step: the resulting population should generalize beyond the discovery data, exhibit the intended functional selectivity, and contribute causally to model behavior. We therefore evaluate SER along these three dimensions. First, frozen SER sets generalize to held-out conversations across four model families. Second, on independent empathy and emotional-support datasets, SER recruitment aligns more closely with affective empathic expression than with the magnitude of user distress. A controlled 2×2 empathy–support factorial experiment further separates empathic acknowledgment from generic support and shows that empathy contributes substantially more to SER recruitment than support alone. Finally, direct activation intervention produces a consistent bidirectional effect: suppressing SER reduces expressed-empathy scores, whereas amplifying SER increases them, with substantially smaller changes in general reasoning and generation capabilities.

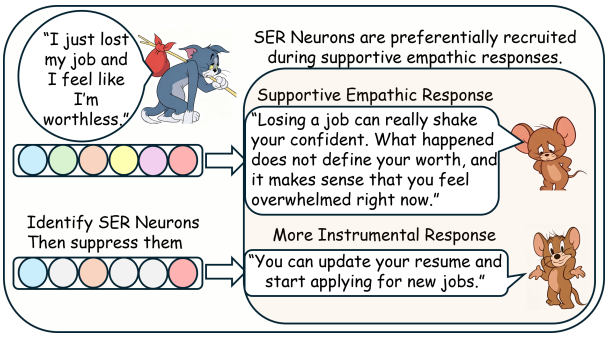


Figure 1: **Example.** With the user context held fixed, suppressing the SER population shifts the model response from a more affectively supportive form toward a more instrumental one.

Together, these findings support the interpretation of SER as a sparse, response-side component of empathic realization rather than a complete controller of emotional-support behavior, and provide evidence for a partial mechanistic dissociation between distress representation and empathic response realization in LLMs.

Contributions. Our work makes three key contributions: **(1) Conceptually,** we formulate distress representation and empathic response realization as mechanistically distinct sub-processes rather than uniform manifestations of a single underlying circuit, thereby providing empirical evidence that they remain only partially coupled inside the model. **(2) Methodologically,** we introduce a counterfactual DiD framework tailored to localizing sparse response-side FFN populations while explicitly suppressing confounding signals from input-side distress sensitivity and generic supportive language. **(3) Empirically,** we demonstrate across four distinct model families that the identified SER populations reliably generalize beyond discovery distributions, preferentially track empathic expression over generic support or input distress magnitude, and causally steer expressed-empathy behavior while largely preserving broader model capabilities.

2 Related Work

Neuron-level mechanistic interpretability. Neuron-level analysis has long been used to localize behaviorally relevant representations in neural language models (Dalvi et al., 2019; Bau et al., 2019), although activation selectivity alone does not establish causal use (Antverg & Belinkov, 2022). Transformer FFNs have subsequently been interpreted as key–value memories or concept-promoting modules (Geva et al., 2021; 2022), with sparse neuron populations linked to factual knowledge (Dai et al., 2022; Meng et al., 2022), multilingual processing (Tang et al., 2024), social bias (Yang et al., 2024; Pan et al., 2026), safety (Chen et al., 2025; Zhao et al., 2025), and cultural understanding (Yamamoto et al., 2026). Recent work increasingly complements localization with causal intervention: shared safety neurons can mediate cross-lingual safety transfer (Zhang et al., 2026), while extremely sparse keystone neurons can exert broad functional effects across tasks (Ji et al., 2026). Complementary studies examine cross-model neuron recurrence (Gurnee et al., 2023; 2024), higher-order neuron coalitions (Chowdhury et al., 2026), and interpretable multilingual circuits assembled from causally localized neurons (Xiao et al., 2026). Our work differs in its target: rather than asking whether an affective property is represented, we isolate a response-side population associated with empathic realization after discounting both input-side distress sensitivity and generic supportive language.

Affective and communicative neurons. Neuron-level affective analysis has identified units associated with generated sentiment (Radford et al., 2017), neurons selective for basic emotion categories (Lee et al., 2025), and neuron populations jointly related to emotion and rhetorical expression (Zheng et al., 2026). These studies show that affective and communicative properties can be localized and manipulated at the neuron level. However, such selectivity does not by itself distinguish whether a neuron encodes the user's affective state, reflects generic supportive style, or contributes specifically to distress-conditioned empathic expression. SER targets this distinction through matched response-side contrasts followed by held-out functional and causal evaluation.

Empathy and emotional-support modeling. Empathy in dialogue systems has primarily been studied behaviorally. EmpatheticDialogues introduced a benchmark for empathetic open-domain dialogue generation (Rashkin et al., 2019); EPITOME decomposes expressed empathy into Emotional Reaction, Interpretation, and Exploration (Sharma et al., 2020); and ESConv models emotional support through explicit support strategies (Liu et al., 2021). Recent LLM-based work further improves empathic response generation through emotional-validation alignment (Son et al., 2026) and emotional-support strategy selection through causal policy learning (Wang et al., 2026). These approaches characterize or optimize empathetic and supportive behavior at the response or dialogue-policy level, but do not localize the internal model components that causally contribute to empathic response realization. We therefore use these behavioral frameworks as external evaluation signals rather than neuron-selection objectives while studying the corresponding response-side mechanism inside LLMs.

3 Method

We study SER in three stages: counterfactual identification, functional characterization of the frozen neuron sets, and causal activation intervention. Throughout the paper, a neuron denotes a signed post-gating FFN dimension at the input to the MLP down-projection.

3.1 SER Identification

Counterfactual supportive–neutral pairs. We use EmpatheticDialogues (Rashkin et al., 2019) and partition user contexts into distress (D) and non-distress control (C) conditions using the emotion labels in Section A. For each context x, we construct matched supportive and neutral responses, y_x^+ and y_x^-, while preserving factual content, topic, and structure. The supportive response retains emotional acknowledgment, validation, reassurance, and interpersonal support, whereas the neutral response remains relevant but removes these elements.

To limit surface-form variation between the two responses, we require

$$0.75 \leq \frac{\text{len}(y_x^+)}{\text{len}(y_x^-)} \leq 1.33, \qquad J(y_x^+, y_x^-) \geq 0.25, \tag{1}$$

where $J(\cdot, \cdot)$ denotes lexical Jaccard similarity over lowercased alphanumeric tokens longer than two characters. The length bounds impose an approximately symmetric multiplicative tolerance around equal response length ($1/0.75 \approx 1.33$), while $J \geq 0.25$ enforces a minimum level of lexical overlap. Both thresholds are fixed pair-quality constraints and are not tuned using downstream neuron-selection or evaluation results. Length and lexical overlap between each rewrite and the original gold response are recorded only as diagnostics and do not affect pair acceptance.

This construction forms a 2×2 design over context (distress vs. control) and response (supportive vs. neutral). Conversations are split into 70% discovery and 30% held-out data with no source overlap.

Difference-in-differences identification. Let $h_{t,i}$ denote the signed activation of neuron i at response token t. Its response-level activation is

$$A_i(x, y) = \frac{1}{|T_y|} \sum_{t \in T_y} h_{t,i}, \tag{2}$$

where T_y is the set of response-token positions. For each matched pair, we first compute the supportive–neutral activation difference and then contrast this effect between distress and control contexts:

$$\begin{aligned} \delta_i(x) &= A_i(x, y_x^+) - A_i(x, y_x^-), \\ \text{DiD}_i &= \mathbb{E}_{x \in D}[\delta_i(x)] - \mathbb{E}_{x \in C}[\delta_i(x)]\,. \end{aligned} \tag{3}$$

The first contrast isolates response-side activation associated with the supportive realization while holding the user context fixed. The second subtracts the same contrast in non-distress contexts, reducing activation attributable to generic supportive language.

We rank neurons using the standardized effect

$$\text{SERScore}_i = \frac{\text{DiD}_i}{s_{p,i}}, \tag{4}$$

where $s_{p,i}$ is the pooled standard deviation of the paired supportive–neutral effects.

Before ranking, neurons are filtered using pre-specified statistical, cross-fold stability, and activation-reliability criteria (Section B.2). Raw significance values are computed with a one-sided normal-approximation Wald test for $H_1 : \text{DiD}_i > 0$ and then corrected across all FFN neurons within each model using Benjamini–Hochberg FDR control (Benjamini & Hochberg, 1995). Eligible neurons are ranked globally by SERScore, and the highest-ranked neurons are retained up to a fixed cardinality equal to 1% of all FFN neurons in the model. We additionally construct layer-matched random, high-activation, and distress-control neuron sets. Their construction and the complete statistical filtering procedure are reported in Section B.2. All neuron sets are frozen before held-out and downstream evaluation.

3.2 Functional Characterization

Without re-selection, we evaluate the frozen neuron sets on independent empathy and emotional-support data. For the first $K = \min(8, |y|)$ response-aligned prediction states, we measure early-response recruitment relative to the prompt-end state:

$$\Delta A_i(x, y) = A_i^{\text{early}}(x, y) - A_i^{\text{prompt}}(x). \tag{5}$$

After out-of-fold normalization, recruitment of a frozen neuron set $\mathcal{S}$ is summarized as

$$\text{SetScore}_{\mathcal{S}}(x) = \frac{1}{|\mathcal{S}|} \sum_{i \in \mathcal{S}} z_i(x), \tag{6}$$

where $z_i(x)$ denotes the cross-fitted normalized recruitment of neuron i.

Affective versus instrumental support. On ESConv (Liu et al., 2021), we group Reflection of Feelings and Affirmation/Reassurance as *affective* responses, and Providing Suggestions and Information as *instrumental* responses. We compare their SER recruitment separately under low- and high-distress contexts and measure whether the affective–instrumental gap changes with distress magnitude:

$$\mathrm{DiD}_{\mathrm{ESConv}} = \big(\mu_{HA} - \mu_{HI}\big) - \big(\mu_{LA} - \mu_{LI}\big), \tag{7}$$

where H/L denote high/low distress and A/I denote affective/instrumental responses. A value near zero indicates that the affective–instrumental recruitment difference is only weakly modulated by distress magnitude.

Expressed-empathy mechanisms. On EPITOME (Sharma et al., 2020), we examine Emotional Reaction, Interpretation, and Exploration separately. For each mechanism $m \in \{\mathrm{ER}, \mathrm{IP}, \mathrm{EX}\}$, we contrast the highest and lowest annotation levels:

$$\Delta_{2-0}^{(m)} = \mathbb{E}[\mathrm{SetScore}_{\mathcal{S}} \mid y_m = 2] - \mathbb{E}[\mathrm{SetScore}_{\mathcal{S}} \mid y_m = 0]\,. \tag{8}$$

Emotional Reaction provides the closest mechanism-level correspondence to affective empathic realization, while Interpretation and Exploration provide complementary comparisons.

Disentangling empathy from generic support. To test whether SER recruitment is driven by empathic realization rather than supportive language alone, we construct matched responses for 1,000 contexts under a 2×2 factorial design: EMPATHY+SUPPORT (A), EMPATHY-ONLY (B), SUPPORT-ONLY (C), and NEITHER (D). The same response quadruplets are evaluated across all model families using their frozen SER sets. We compare the effect of adding empathy while holding support present ($A-C$) with the effect of adding support while holding empathy present ($A-B$), and additionally report the standard empathy and support main effects and their interaction. Confidence intervals are obtained by conversation-cluster bootstrap over the 1,000 contexts.

Together, these analyses test whether frozen SER recruitment is more closely associated with empathic response realization than with either user-distress magnitude or generic supportive language.

3.3 CAUSAL INTERVENTION

We test the causal contribution of frozen SER by scaling its post-gating FFN activations while leaving all other neurons unchanged:

$$\widetilde{h}_{t,i}^{(l)} = \begin{cases} \alpha\, h_{t,i}^{(l)}, & (l,i) \in \mathcal{S}, \\ h_{t,i}^{(l)}, & (l,i) \notin \mathcal{S}, \end{cases} \qquad \alpha \in \{0, 0.5, 1, 1.5, 2\}. \tag{9}$$

The unmodified model corresponds to $\alpha = 1$; $\alpha < 1$ suppresses the selected neurons, whereas $\alpha > 1$ amplifies them. For each intervention strength, we report the paired mean change relative to the unmodified model:

$$\Delta(\alpha) = \frac{1}{N} \sum_{x=1}^{N} [Y_x(\alpha) - Y_x(1)]\,. \tag{10}$$

We evaluate whether intervention produces a consistent bidirectional effect on expressed empathy, compare the same intervention with matched neuron controls, and examine broader emotion-related, reasoning, and generation capabilities to assess functional specificity. The same operator is used for the Top-$p\%$ and depth-restricted ablations in Section 5.5.

4 EXPERIMENTAL SETUP

Models. Our primary model is Qwen2.5-7B-Instruct (Qwen et al., 2025). We independently apply the same identification, functional characterization, and intervention procedures to OLMo-7B-Instruct (Groeneveld et al., 2024), Mistral-7B-Instruct-v0.3 (Jiang et al., 2023), and ChatGLM3-6B (Team GLM et al., 2024). Qwen2.5-32B-Instruct is additionally used to examine depth organization across model scales. SER neurons are identified independently within each architecture; no neurons are transferred or matched across models.

Table 1: **Held-out validation of frozen neuron sets.** Controls are matched to SER in neuron count and layer distribution. Higher is better for all metrics.

	Held-out DiD ↑				Supportive–Neutral AUC ↑				Pair Accuracy ↑			
Model	**SER**	Rand.	High-Act.	Dist.	**SER**	Rand.	High-Act.	Dist.	**SER**	Rand.	High-Act.	Dist.
Qwen2.5-7B	**.0137**	-.0003	-.0016	.0034	**.6217**	.4377	.5057	.4977	**.8667**	.3132	.4846	.5277
OLMo-7B	**.0204**	-.0003	-.0037	.0015	**.5987**	.4146	.3746	.4764	**.9350**	.3050	.1950	.4100
Mistral-7B-v0.3	**.0111**	-.0001	-.0004	.0006	**.6531**	.4263	.2307	.4675	**.9319**	.3345	.1102	.3692
ChatGLM3-6B	**.0136**	-.0002	-.0020	.0010	**.6536**	.4711	.4375	.4930	**.9020**	.4211	.3247	.5007

Note. Rand.: layer-matched random neurons; High-Act.: layer-matched high-activation neurons; Dist.: layer-matched distress-control neurons.

Datasets and Benchmarks. EmpatheticDialogues (Rashkin et al., 2019) is used for counterfactual construction and SER identification; ESConv (Liu et al., 2021) and EPITOME (Sharma et al., 2020) provide independent support-strategy and expressed-empathy annotations. For intervention scoring, we follow the released EPITOME RoBERTa-base bi-encoder implementation and use separate mechanism-specific Reddit checkpoints for Emotional Reaction, Interpretation, and Exploration; exact checkpoint names and scoring details are given in Section D.1. Broader emotion-related evaluation uses EmoBench (Sabour et al., 2024) and EQ-Bench (Paech, 2023). General capabilities are measured with ARC-Challenge (Clark et al., 2018) and MMLU (Hendrycks et al., 2021); BLEU (Papineni et al., 2002) and perplexity provide generation-level controls.

Inference Configuration. All models use their released chat templates and bfloat16 inference. Counterfactual identification, functional characterization, and intervention use maximum input budgets of 672, 768, and 1,024 tokens, respectively. Unless otherwise specified, generation uses greedy decoding, with at most 256 new tokens for generative tasks and 16 for short-answer tasks. Additional implementation details are provided in Section E.

5 Results

5.1 SER Organization and Generalization

Frozen SER generalizes to unseen conversations. Table 1 evaluates the frozen neuron sets on held-out conversations. Across all four model families, SER retains a positive held-out DiD, whereas the random and high-activation controls remain near zero or negative and the distress-control set is consistently weaker. The same separation appears in supportive–neutral discrimination: SER achieves AUC values of .599–.654 and pair accuracies of .867–.935, indicating that aggregate SER activation is typically higher for supportive responses than for their matched neutral counterfactuals. Together, these results show that the identified representation transfers beyond the discovery data.

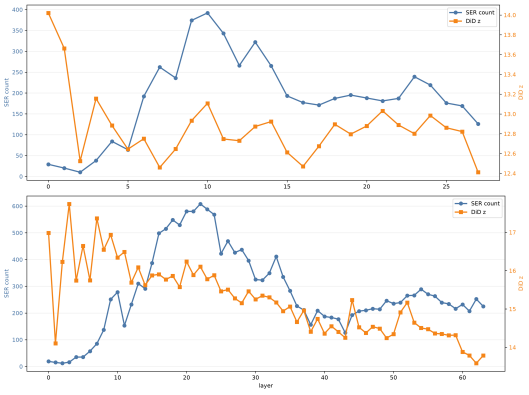


Figure 2: **Depth-wise organization of SER.** Layer-wise SER count and mean standardized DiD for Qwen2.5-7B (top) and Qwen2.5-32B (bottom).

SER exhibits structured depth-wise organization. As shown in Figure 2, SER neurons are sparse in early layers, become more concentrated at intermediate depths, and remain distributed across later layers in both Qwen2.5-7B and Qwen2.5-32B. The layer-wise neuron count and mean standardized DiD follow different profiles, separating *SER density* from *SER selectivity*. The other 7B level models show related but architecture-dependent depth profiles (Figure 5 and Section B.5).

Table 2: **Functional characterization of frozen neuron sets on EPITOME and ESConv.** SER shows increasing recruitment with Emotional Reaction and consistently higher recruitment for affective than instrumental support across all four model families.

		EPITOME			ESConv: Low Distress		ESConv: High Distress	
Model	Set	ER0	ER1	ER2	Inst.	Aff.	Inst.	Aff.
Qwen2.5-7B	**SER**	**-.0377**	**+.0395**	**+.2778**	**-.0813**	**+.1105**	**-.0945**	**+.0944**
	Rand.	.0000	+.0008	-.0037	-.0007	-.0005	+.0005	+.0007
	High-Act.	+.0012	-.0017	-.0063	+.0029	-.0048	+.0029	-.0040
	Dist.-Ctrl.	+.0039	-.0072	-.0108	+.0073	+.0031	-.0115	-.0219
OLMo-7B	**SER**	**-.0232**	**+.0204**	**+.1932**	**-.0415**	**+.0574**	**-.0509**	**+.0555**
	Rand.	-.0002	+.0014	-.0057	-.0008	.0000	-.0004	+.0004
	High-Act.	+.0023	-.0009	-.0263	+.0035	-.0047	+.0029	-.0043
	Dist.-Ctrl.	+.0088	-.0169	-.0207	+.0126	+.0035	-.0114	-.0243
Mistral-7B	**SER**	**-.0202**	**+.0120**	**+.2016**	**-.0365**	**+.0505**	**-.0422**	**+.0462**
	Rand.	+.0003	-.0007	+.0004	-.0005	-.0005	+.0009	.0000
	High-Act.	+.0023	-.0032	-.0131	+.0040	-.0036	+.0034	-.0057
	Dist.-Ctrl.	+.0058	-.0108	-.0155	+.0130	-.0049	-.0080	-.0390
ChatGLM3-6B	**SER**	**-.0330**	**+.0407**	**+.2086**	**-.0627**	**+.0846**	**-.0601**	**+.0872**
	Rand.	-.0005	+.0004	+.0034	-.0002	-.0018	+.0022	+.0010
	High-Act.	+.0033	-.0044	-.0191	+.0064	-.0083	+.0075	-.0088
	Dist.-Ctrl.	+.0004	-.0026	+.0088	-.0006	+.0010	+.0019	+.0045

Note. ER0–ER2: Emotional Reaction levels 0–2; Inst./Aff.: instrumental/affective support; Rand.: layer-matched random; High-Act.: layer-matched high-activation; Dist.-Ctrl.: layer-matched distress-control. Values are mean cross-fitted normalized recruitment scores.

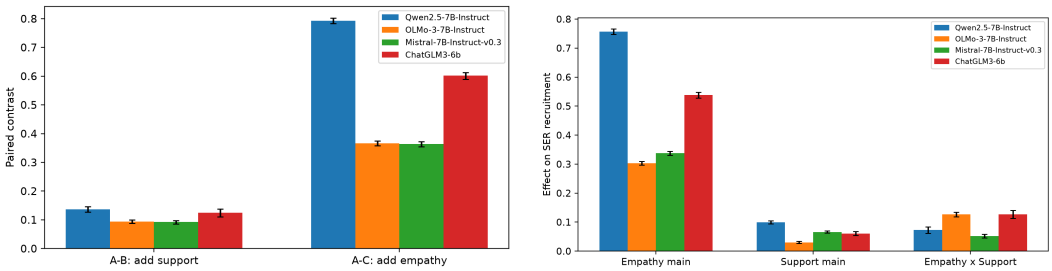
5.2 Functional Selectivity of SER

SER recruitment increases with expressed Emotional Reaction. We next characterize what frozen SER recruitment reflects using independent EPITOME and ESConv annotations. Across both datasets, recruitment aligns more closely with affective empathic realization than with user-distress magnitude (Table 2). Specifically, on EPITOME, SER recruitment increases monotonically from ER0 to ER2 across all four model families, with ER2–ER0 differences of 0.3155 (Qwen2.5), 0.2164 (OLMo), 0.2218 (Mistral), and 0.2416 (ChatGLM3). Matched random, high-activation, and distress-control sets do not reproduce this trend. Because SER is frozen prior to EPITOME evaluation, this association is independent of the discovery contrast.

SER preferentially tracks affective rather than instrumental support. On ESConv, SER recruitment is consistently higher for affective than instrumental responses under both low- and high-distress conditions (Table 2). Furthermore, the affective–instrumental gap changes little with distress level: moderation effects remain close to zero across all four models, ranging from -0.0029 to $+0.0075$. Matched controls do not reproduce this separation, supporting an association with response-side affective realization rather than distress magnitude.

5.3 Disentangling Empathy and Support

SER recruitment is driven primarily by empathy rather than generic support. As shown in Figure 3, adding empathy while holding support present ($A - C$) produces substantially larger SER recruitment than adding support while holding empathy present ($A - B$) across all four model families. The full factorial analysis shows the same asymmetry: the empathy main effect is 0.303–0.757, versus only 0.030–0.100 for support, while the empathy$\times$support interaction remains smaller (0.052–0.127). Together, these results argue against a simple generic-support explanation and indicate that empathy is the dominant driver of SER recruitment, with support providing a weaker secondary modulation.



(a) **Paired contrasts.** (b) **Factorial effects.**

Figure 3: **SER recruitment is driven primarily by empathy rather than generic support.** Results from the 2×2 factorial response-control experiment over 1,000 ESConv contexts. **(a)** Paired effects of adding support ($A - B$) or empathy ($A - C$). **(b)** Empathy and support main effects and their interaction. Error bars denote 95% conversation-cluster bootstrap confidence intervals.

5.4 Causal Intervention on SER

SER causally modulates expressed empathy. Scaling frozen SER activations produces a consistent bidirectional effect on EPITOME across four model families (Table 3). Complete suppression reduces expressed-empathy scores by 10.3–31.9%, and partial suppression is negative in every model. Conversely, amplification consistently increases EPITOME, with $\alpha = 2$ yielding gains of 1.5–9.3%. Thus, SER activity reproducibly contributes to expressed empathy, with suppression producing larger effects than amplification beyond the native operating level.

The effect is functionally selective. Under these interventions, general reasoning and generation benchmarks change substantially less, while broader emotion-related metrics show smaller or directionally mixed effects (Table 3). Suppression of high-activation control neurons can reduce EPITOME in some models, but these controls do not reproduce the consistent cross-model suppression–amplification pattern of SER (Section D.3). The causal evidence rests on bidirectionality, cross-model consistency, and functional specificity rather than suppression magnitude alone.

SER is a component, not a complete controller, of empathic realization. The suppression–amplification asymmetry suggests that SER contributes to response-side empathic realization but is insufficient to determine the full quality of an emotional-support response. Increasing SER activity cannot by itself improve other required computations, such as situational appraisal, affective understanding, strategy selection, and contextual integration. Together, these results position SER as a causal component of empathic response realization rather than a stand-alone controller of the complete support process.

5.5 Top-p% Ablations

Sparse top-ranked neurons already produce substantial causal effects. As shown in Table 4, masking FFN neurons by SERScore reduces EPITOME-based expressed empathy at every tested fraction across all four model families. The effect is substantial even at the top 0.1%, ranging from -0.058 for ChatGLM3 to -0.141 for Mistral. With expanding masks, Qwen shows larger drops reaching -0.255 at 3%, while Mistral plateaus near its 0.2% level across wider masks. OLMo and ChatGLM3 exhibit smaller yet persistent reductions throughout the tested range. Thus, causal effects emerge within a sparse top-ranked subset and extend beyond the fixed Top-1% SER threshold, though their dependence on p is model-specific and non-monotonic.

6 Limitations

Construct validity. Counterfactual and factorial controls reduce major confounds but cannot remove all differences in lexical realization, discourse style, or pragmatic naturalness; both empathy and support remain operationalized through constructed responses.

Table 3: **Behavioral effects of scaling frozen SER activations.** Baseline rows report absolute scores at $\alpha = 1$; all intervention rows report paired mean changes from the corresponding baseline. ESConv columns report affective- and instrumental-support probabilities at baseline and probability changes under intervention.

		General Reasoning		Generation		Emotion		Empathy		
Model	Scale	ARC-C	MMLU	BLEU	PPL ↓	EmoBench	EQ-Bench	EPITOME	ESConv-Aff.	ESConv-Inst.
Qwen2.5-7B	Baseline ($\alpha = 1$)	.898	.708	.320	2.218	.575	.659	.678	.411	.358
	$\alpha = 0$	-.017	-.026	-.022	+.018	-.035	+.019	**-.216**	**-.049**	**+.050**
	$\alpha = .5$	-.001	-.003	-.003	+.004	+.003	+.003	**-.121**	**-.026**	**+.023**
	$\alpha = 1.5$	.000	-.004	.000	+.004	-.007	-.003	**+.035**	**+.030**	**-.029**
	$\alpha = 2$	-.007	-.011	-.002	+.017	-.018	+.027	**+.042**	**+.058**	**-.050**
Mistral-7B	Baseline ($\alpha = 1$)	.765	.585	.241	1.956	.493	.624	.639	.313	.486
	$\alpha = 0$	-.009	-.004	-.017	+.016	.000	-.007	**-.168**	**-.058**	**+.041**
	$\alpha = .5$	+.001	.000	-.007	+.004	-.003	-.001	**-.066**	**-.035**	**+.026**
	$\alpha = 1.5$	-.005	.000	+.005	+.002	-.015	-.020	**+.016**	**+.042**	**-.034**
	$\alpha = 2$	-.018	-.002	+.006	+.011	-.022	-.029	**+.023**	**+.088**	**-.068**
OLMo-7B	Baseline ($\alpha = 1$)	.776	.575	.280	2.896	.458	.681	.715	.444	.330
	$\alpha = 0$	-.003	-.005	-.003	+.055	-.003	-.027	**-.074**	**-.097**	**+.099**
	$\alpha = .5$	-.003	-.002	.000	+.009	+.005	+.002	**-.019**	**-.044**	**+.042**
	$\alpha = 1.5$	+.003	-.002	-.002	+.005	-.003	+.018	**+.004**	**+.022**	**-.026**
	$\alpha = 2$	+.006	.000	-.004	+.014	-.013	+.003	**+.011**	**+.039**	**-.049**
ChatGLM3-6B	Baseline ($\alpha = 1$)	.568	.444	.170	3.621	.375	.478	.603	.393	.264
	$\alpha = 0$	-.009	-.008	+.005	+.011	+.020	+.002	**-.109**	**-.021**	**+.016**
	$\alpha = .5$	-.001	-.004	+.005	+.003	+.015	-.014	**-.048**	**-.009**	**+.008**
	$\alpha = 1.5$	+.007	+.001	-.006	+.003	+.003	+.008	**+.031**	**+.008**	**-.020**
	$\alpha = 2$	+.002	-.003	-.012	+.010	+.003	-.010	**+.056**	**+.016**	**-.046**

Note. Baseline rows report absolute values; intervention rows report paired differences from the corresponding $\alpha = 1$ baseline. Positive ΔPPL indicates worse language-modeling performance. EPITOME is the mechanism-proximal expressed-empathy endpoint. ESConv-Aff. is the summed probability of *Reflection of Feelings* and *Affirmation/Reassurance*; ESConv-Inst. is the summed probability of *Providing Suggestions* and *Information*. ESConv values measure strategy composition rather than overall response quality, so larger values are not uniformly better.

Table 4: **Top-$p\%$ ablation under the** SERScore **ranking.** Baseline reports the absolute EPITOME score under the separately retrained RoBERTa evaluator used for this extended ablation analysis. All remaining entries report paired ΔEPITOME after masking the highest-ranked FFN neurons at each fraction p. All intervention effects have 95% bootstrap confidence intervals below zero ($N = 3023$ per model).

Model	Baseline	0.1%	0.2%	0.3%	0.4%	0.5%	0.6%	0.7%	0.8%	0.9%	1.0%	2.0%	3.0%	4.0%	5.0%
Qwen2.5-7B	.678	-.140	-.139	-.141	-.123	-.145	-.149	-.143	-.153	-.159	-.214	-.231	-.255	-.248	-.239
Mistral-7B	.665	-.141	-.211	-.207	-.215	-.221	-.209	-.216	-.215	-.206	-.213	-.190	-.193	-.211	-.217
OLMo-7B	.817	-.080	-.069	-.070	-.070	-.083	-.074	-.096	-.113	-.104	-.107	-.083	-.105	-.109	-.093
ChatGLM3-6B	.655	-.058	-.058	-.051	-.054	-.057	-.081	-.084	-.088	-.098	-.089	-.099	-.095	-.104	-.098

Generalizability. Our investigation focuses primarily on English emotional-support data and open-weight models within the 6B–32B parameter range. Consequently, whether the identified response-side functional organization robustly transfers across broader linguistic contexts, extreme model scales, closed access regimes, and multimodal interaction settings remains an open question for future work.

7 Conclusion

We identify SER, a sparse response-side FFN population that contributes to empathic response realization. Across model families, SER generalizes beyond discovery data, is preferentially recruited during affective empathic responses, and exhibits reproducible bidirectional effects on both expressed empathy and support-strategy realization under activation intervention. Together, these findings support a partial mechanistic dissociation between representing user distress and realizing an empathic response in LLMs.

AI USE STATEMENT

Generative AI tools assisted with language editing, manuscript organization, code review and debugging, and identifying potential inconsistencies in experimental outputs. All experimental designs, implementation decisions, data processing, statistical analyses, reported results, and scientific conclusions were independently verified by the authors, who take full responsibility for the paper.

ETHICS STATEMENT

This work studies internal mechanisms of emotional-support generation and makes no clinical, diagnostic, or therapeutic claims. We use publicly released research datasets under their stated research conditions and collect no new human-subject data. Activation interventions are mechanistic probes rather than deployment recommendations. Because emotionally supportive systems may be used in sensitive contexts, our findings should not be interpreted as establishing clinical efficacy, safety, or suitability for high-stakes mental-health applications.

REPRODUCIBILITY STATEMENT

The main text specifies the activation site, counterfactual construction, selection statistic, significance test, functional analysis, intervention operator, model families, context budgets, and evaluation endpoints. The appendices provide exact emotion-label sets, statistical filtering thresholds, cross-fitting, EPITOME classifier checkpoints and scoring definitions, discovery diagnostics, and experiment settings. Experiment code, frozen neuron-set artifacts, and sample-level intervention outputs will be included in the anonymized supplementary material.

REFERENCES

Omer Antverg and Yonatan Belinkov. On the pitfalls of analyzing individual neurons in language models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=8uz0EWPQIMu.

Anthony Bau, Yonatan Belinkov, Hassan Sajjad, Nadir Durrani, Fahim Dalvi, and James Glass. Identifying and controlling important neurons in neural machine translation. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=H1z-PsR5KX.

Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1):289–300, 1995. doi: 10.1111/j.2517-6161.1995.tb02031.x.

Jianhui Chen, Xiaozhi Wang, Zijun Yao, Yushi Bai, Lei Hou, and Juanzi Li. Towards understanding safety alignment: A mechanistic perspective from safety neurons. In *Advances in Neural Information Processing Systems*, volume 38, pp. 12665–12700. Curran Associates, Inc., 2025. doi: 10.52202/085713-0426. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/12a00d85a76fe258e1242c3aced03250-Paper-Conference.pdf.

Tanya Chowdhury, Atharva Nijasure, Yair Zick, and James Allan. Hedonic neurons: A mechanistic mapping of latent coalitions in transformer MLPs. In *International Conference on Learning Representations*, 2026.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try ARC, the AI2 reasoning challenge. *arXiv preprint arXiv:1803.05457*, 2018.

Damai Dai, Li Dong, Yaru Hao, Zhifang Sui, Baobao Chang, and Furu Wei. Knowledge neurons in pretrained transformers. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8493–8502. Association for Computational Linguistics, 2022. doi: 10.18653/v1/2022.acl-long.581.

Fahim Dalvi, Nadir Durrani, Hassan Sajjad, Yonatan Belinkov, Anthony Bau, and James R. Glass. What is one grain of sand in the desert? analyzing individual neurons in deep NLP models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 6309–6317, 2019. doi: 10.1609/aaai.v33i01.33016309.

Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. Transformer feed-forward layers are key-value memories. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 5484–5495. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.emnlp-main.446.

Mor Geva, Avi Caciularu, Kevin Ro Wang, and Yoav Goldberg. Transformer feed-forward layers build predictions by promoting concepts in the vocabulary space. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 30–45. Association for Computational Linguistics, 2022. doi: 10.18653/v1/2022.emnlp-main.3.

Dirk Groeneveld, Iz Beltagy, Evan Walsh, Akshita Bhagia, Rodney Kinney, Oyvind Tafjord, Ananya Jha, Hamish Ivison, Ian Magnusson, Yizhong Wang, et al. OLMo: Accelerating the science of language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15789–15809. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.acl-long.841.

Wes Gurnee, Neel Nanda, Matthew Pauly, Katherine Harvey, Dmitrii Troitskii, and Dimitris Bertsimas. Finding neurons in a haystack: Case studies with sparse probing. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856. URL https://openreview.net/forum?id=JYs1R9IMJr.

Wes Gurnee, Theo Horsley, Zifan Carl Guo, Tara Rezaei Kheirkhah, Qinyi Sun, Will Hathaway, Neel Nanda, and Dimitris Bertsimas. Universal neurons in GPT2 language models. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL https://openreview.net/forum?id=ZeI104QZ8I.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *International Conference on Learning Representations*, 2021. URL https://mlanthology.org/iclr/2021/hendrycks2021iclr-measuring/.

Xiangtian Ji, Yuxin Chen, Zhengzhou Cai, Xiang Wang, An Zhang, and Tat-Seng Chua. Tiny brains, giant impact: Uncovering the keystone neurons of LLM with just a few prompts. In *Forty-Third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=vbKUSHz68Z.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7B. *arXiv preprint arXiv:2310.06825*, 2023. doi: 10.48550/arXiv.2310.06825.

Richard S. Lazarus. Progress on a cognitive-motivational-relational theory of emotion. *American Psychologist*, 46(8):819–834, 1991. doi: 10.1037/0003-066X.46.8.819.

Jaewook Lee, Woojin Lee, Oh-Woog Kwon, and Harksoo Kim. Do large language models have "emotion neurons"? investigating the existence and role. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 15617–15639. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.findings-acl.806.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Inference-time intervention: Eliciting truthful answers from a language model. In *Advances in Neural Information Processing Systems*, volume 36, pp. 41451–41530, 2023.

Siyang Liu, Chujie Zheng, Orianna Demasi, Sahand Sabour, Yu Li, Zhou Yu, Yong Jiang, and Minlie Huang. Towards emotional support dialog systems. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 3469–3483. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.acl-long.269.

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

Samuel J. Paech. EQ-Bench: An emotional intelligence benchmark for large language models. *arXiv preprint arXiv:2312.06281*, 2023.

Jinhao Pan, Chahat Raj, Anjishnu Mukherjee, Sina Mansouri, Bowen Wei, Shloka Yada, and Ziwei Zhu. Knowing bias, doing better: Mitigating social bias in LLMs via know-bias neuron enhancement. In *Forty-Third International Conference on Machine Learning*, 2026.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pp. 311–318. Association for Computational Linguistics, 2002. doi: 10.3115/1073083.1073135.

Qwen, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report, 2025. URL https://arxiv.org/abs/2412.15115. arXiv:2412.15115.

Alec Radford, Rafal Jozefowicz, and Ilya Sutskever. Learning to generate reviews and discovering sentiment. *arXiv preprint arXiv:1704.01444*, 2017.

Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. Towards empathetic open-domain conversation models: A new benchmark and dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 5370–5381. Association for Computational Linguistics, 2019. doi: 10.18653/v1/P19-1534.

Sahand Sabour, Siyang Liu, Zheyuan Zhang, June Liu, Jinfeng Zhou, Alvionna Sunaryo, Tatia Lee, Rada Mihalcea, and Minlie Huang. EmoBench: Evaluating the emotional intelligence of large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 5986–6004. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.acl-long.326.

Ashish Sharma, Adam Miner, David Atkins, and Tim Althoff. A computational approach to understanding empathy expressed in text-based mental health support. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 5263–5276. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.emnlp-main.425.

Suhyune Son, Jungwoo Lim, Myunghoon Kang, Seongtae Hong, Yuna Hur, Evelyn H. Zi, and Heuiseok Lim. I don't need solution. I need emotional support: Empathetic LLMs based on emotional validation. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 1–26. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.findings-acl.1.

Tianyi Tang, Wenyang Luo, Haoyang Huang, Dongdong Zhang, Xiaolei Wang, Wayne Xin Zhao, Furu Wei, and Ji-Rong Wen. Language-specific neurons: The key to multilingual capabilities in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 5701–5715. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.acl-long.309.

Team GLM, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Dan Zhang, Diego Rojas, Guanyu Feng, Hanlin Zhao, et al. ChatGLM: A family of large language models from GLM-130B to GLM-4 all tools. *arXiv preprint arXiv:2406.12793*, 2024.

Xv Wang, Wang Zhenyu, Guanyu Zheng, and Rui Zhang. Causal-ESC: Reliable policy learning for emotional support conversation via causal inference. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 19386–19401. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.acl-long.886.

Theresa Wiseman. A concept analysis of empathy. *Journal of Advanced Nursing*, 23(6):1162–1167, 1996. doi: 10.1046/j.1365-2648.1996.12213.x.

Yuxin Xiao, Zhen Huang, Wenxiao Wang, Yan Zhao, Zhihong Gu, Binbin Lin, Xiaofei He, Xu Shen, and Jieping Ye. How do language models speak languages? a case study on unintended code-switching. In *Forty-Third International Conference on Machine Learning*, 2026.

Taisei Yamamoto, Ryoma Kumon, Danushka Bollegala, and Hitomi Yanaka. Neuron-level analysis of cultural understanding in large language models. In *International Conference on Learning Representations*, 2026.

Nakyeong Yang, Taegwan Kang, Stanley Jungkyu Choi, Honglak Lee, and Kyomin Jung. Mitigating biases for instruction-following language models via bias neurons elimination. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 9061–9073. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.acl-long.490.

Xianhui Zhang, Chengyu Xie, Linxia Zhu, Yonghui Yang, Weixiang Zhao, Zifeng Cheng, Cong Wang, Fei Shen, and Tat-Seng Chua. Who transfers safety? identifying and targeting cross-lingual shared safety neurons. In *Forty-Third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=mNX4kpc371.

Yiran Zhao, Wenxuan Zhang, Yuxi Xie, Anirudh Goyal, Kenji Kawaguchi, and Michael Qizhe Shieh. Understanding and enhancing safety mechanisms of LLMs via safety-specific neuron. In *International Conference on Learning Representations*, 2025.

Li Zheng, Xin Zhang, Shuyi He, Fei Li, Chong Teng, Jiang-Ming Yang, Donghong Ji, and Zhuang Li. Are emotion and rhetoric neurons in LLM? neuron recognition and adaptive masking for emotion-rhetoric prediction steering. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 4206–4216. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.acl-long.193.

A Data Construction and Counterfactual Design

A.1 Operational Definitions and Emotion Conditions

Emotion conditions. The distress-label set is

```
afraid, angry, annoyed, anxious, apprehensive, ashamed,
devastated, disappointed, disgusted, embarrassed, furious,
guilty, jealous, lonely, sad, terrified.
```

The non-distress control-label set is

```
prepared, proud, excited, joyful, grateful, hopeful,
confident, caring, trusting, faithful, content, impressed,
surprised, anticipating.
```

Samples outside these two operational label sets are not included in the identification contrast.

A.2 Minimal-Pair Counterfactual Rewriting

For each original assistant response, we construct a supportive–neutral minimal pair matched in factual content, subject, sentence count, and lexical realization. The supportive response preserves emotional acknowledgment, validation, reassurance, and interpersonal support; the neutral response remains polite, coherent, relevant, and safe while removing these components. Pairs that fail the length-ratio or lexical-overlap constraints in Equation (1) are regenerated or excluded. These constraints measure pair quality, not empathy.

A.3 COUNTERFACTUAL GENERATION AND VALIDATION

Rewriting model and decoding. For each model-specific identification run, counterfactuals are generated by the same model whose activations are subsequently analyzed. For the primary Qwen2.5-7B-Instruct analysis, generation uses bfloat16 inference with seed 42, the model's released chat template, greedy decoding (`do_sample=False`), at most 256 new tokens, and batch size 16. Each pair carries the protocol version `ciser_minimal_pair_v3`; cached pairs with a mismatched version are discarded and regenerated to prevent resumed runs from mixing construction protocols.

Repair retries. A pair that fails validation is regenerated using a failure-specific repair instruction, including length matching, increased lexical reuse, or a strict-JSON reminder. We allow at most two repair attempts after the initial generation, for three generation attempts in total. Pairs that still fail validation are recorded as `max_retries_exhausted` and excluded.

Acceptance statistics. For the Qwen2.5-7B-Instruct generation run, 53,699 candidate pairs are processed, of which 32,844 pass validation and 20,855 are excluded, corresponding to a validation pass rate of 61.2%. Exclusions comprise 11,362 length-ratio failures, 8,823 supportive–neutral lexical-overlap failures, 549 JSON-parsing failures, and 121 identical-response pairs. These constraints assess pair quality rather than empathy.

Seed-sample filtering and splitting. Adjacent-turn pairs from EmpatheticDialogues `train.jsonl` are retained only if both the user turn and gold response contain at least three words and the two turns come from different speakers. Samples are split at the conversation-source level into 70% discovery and 30% held-out partitions using seed 42, with zero source overlap. The resulting data contain 23,022 discovery samples (11,745 distress and 11,277 control; 10,849 sources) and 9,822 held-out samples (4,959 distress and 4,863 control; 4,650 sources). Cross-validation folds are likewise grouped by source.

A.4 COUNTERFACTUAL REWRITING PROMPT

The following instruction constrains each supportive–neutral pair to preserve content and structure while differing primarily in supportive or empathic wording.

```
Create two minimally different rewrites of the original assistant
↪  reply.

Primary constraint: supportive_response and neutral_response must be a
↪  minimal
pair. They should describe the same facts, same subject, same number
↪  of
sentences, nearly identical length, high lexical/content overlap, and
↪  no new
facts. Differences should be limited to supportive/empathic wording.

The original assistant reply is a factual anchor, not a strict length
↪  template;
do not delete a pair only because it is shorter or longer than the
↪  original.

supportive_response: preserve natural emotional understanding,
↪  interpersonal
support, and an appropriate supportive tone.

neutral_response: remain polite, coherent, relevant, and factually
↪  similar,
but remove emotional validation, empathic acknowledgment, reassurance,
↪  and
interpersonal support. It must not be rude, dismissive, unsafe, or
↪  refusing.

Return only strict JSON:
```

```
{"supportive_response": "...", "neutral_response": "..."}

User:
{user_text}

Original assistant reply:
{gold_response}
```

B SER Identification and Selection Diagnostics

B.1 Identification Activation Extraction

Identification uses a total input budget of 672 tokens. The user-context prefix and candidate response are tokenized separately before concatenation to preserve the response-token boundary used for activation pooling. When the combined sequence exceeds the token budget, user-prefix tokens are removed first while retaining at least one prefix token; any remaining overflow is removed from the response tail. Sequences are left-padded within each batch.

Activations are read at the input to each FFN down-projection, corresponding to the post-gating intermediate representation used throughout the paper. For response-mode extraction, activations are averaged over response-token positions. For prompt-mode extraction, the readout is taken at the final non-padding prompt position.

B.2 Statistical Filtering and Matched Control Construction

Statistical filtering. Discovery uses five source-grouped folds. For each neuron, let $\bar{\delta}_{D,i}$ and $\bar{\delta}_{C,i}$ denote the mean supportive–neutral effects in distress and control contexts, with sample variances $s^2_{D,i}$ and $s^2_{C,i}$ and sample counts n_D and n_C. We test the pre-specified one-sided alternative

$$H_1 : \mathrm{DiD}_i > 0 \tag{11}$$

using the unequal-variance standard error

$$\mathrm{SE}(\mathrm{DiD}_i) = \sqrt{\frac{s^2_{D,i}}{n_D} + \frac{s^2_{C,i}}{n_C}}. \tag{12}$$

The Wald statistic is

$$Z_i = \frac{\mathrm{DiD}_i}{\mathrm{SE}(\mathrm{DiD}_i)}, \qquad p_i = 1 - \Phi(Z_i), \tag{13}$$

where Φ is the standard-normal CDF. Thus, the default significance test is a one-sided normal-approximation Wald test, rather than a permutation test. Benjamini–Hochberg correction (Benjamini & Hochberg, 1995) is then applied to the raw p_i values across all FFN neurons within each model. The resulting adjusted values are denoted q_i.

Eligibility requires:

- a positive supportive–neutral effect in distress contexts;
- a positive distress-conditioned DiD;
- direction consistency of at least 0.60;
- positive DiD in at least 80% of grouped folds;
- Benjamini–Hochberg-adjusted $q \leq 0.05$; and
- mean activation at or above the 20th-percentile reliability threshold.

Eligible neurons are ranked globally by $\mathrm{SERScore}$, and we retain the highest-ranked neurons up to a fixed cardinality equal to 1% of all FFN neurons in the model. The resulting SER set is frozen before all downstream evaluation.

A one-sided sign-flip Monte Carlo permutation test is implemented as an optional diagnostic for selected candidates, but its p-values are not used for the default BH filtering or neuron selection.

Table 5: **Cumulative SER selection funnel across model families.** Each column reports the number of neurons remaining after cumulatively applying the corresponding selection criterion. "Direction" denotes the direction-consistency requirement, "Fold stability" requires a positive DiD in at least 80% of the source-grouped folds, and "BH" denotes Benjamini–Hochberg correction with $q \leq 0.05$. The final column gives the frozen SER set, whose cardinality is fixed to 1% of all FFN neurons in the corresponding model.

Model	All	+ Distress gap	+ Positive DiD	+ Direction consistency	+ Fold stability	+ BH correction	+ Activation reliability	Final SER
Qwen2.5-7B-Instruct	530,432	265,384	167,005	40,463	36,018	31,311	25,203	5,305
OLMo-7B-Instruct	352,256	176,292	114,786	41,326	33,493	24,156	18,541	3,523
Mistral-7B-Instruct-v0.3	458,752	229,707	129,855	57,535	51,402	45,720	36,313	4,588
ChatGLM3-6B	383,488	191,568	125,803	13,292	12,189	10,983	8,369	3,835

Matched control construction. For every layer containing m_l selected SER neurons, each control set contains exactly m_l non-SER neurons. The layer-matched random control is sampled within the corresponding layer. The high-activation control selects neurons with the largest overall mean activation. The distress-control set selects neurons with the largest prompt-level distress–control activation difference. All control sets are non-overlapping with SER and reproduce its complete layer histogram.

B.3 Cross-Model SER Selection Funnel

Table 5 reports the cumulative number of FFN neurons remaining after each selection stage in each model.

Filtering sharply reduces the eligible pool in every model, with the direction-consistency criterion producing the largest drop, especially for ChatGLM3-6B. The final frozen SER sets contain $5{,}305$, $3{,}523$, $4{,}588$, and $3{,}835$ neurons for Qwen2.5-7B-Instruct, OLMo-7B-Instruct, Mistral-7B-Instruct-v0.3, and ChatGLM3-6B, respectively.

B.4 Discovery-Stage Distributional Diagnostics

Discovery distributions. Figure 4 visualizes the successive stages of the selection procedure. The complete FFN population contains a broad range of positive and negative distress-conditioned effects. Eligibility filtering restricts this population toward positive and stable neurons, and the final Top-1% SER set occupies the strongest positive region of the standardized DiD distribution.

The same qualitative structure appears across all four model families. Because this separation is partly induced by the selection procedure itself, we do not use the distributional shift as evidence of out-of-sample validity. The held-out results in Table 1 provide the corresponding independent test.

B.5 Cross-Model Layer-Wise Distribution of SER

To examine whether the depth organization of SER is specific to a single architecture, we visualize the layer-wise distribution of the frozen SER sets across all four model families. For each layer, we report both the number of selected SER neurons and their mean standardized DiD score.

SER exhibits a broadly shared depth organization across architectures. Figure 5 shows that the layer-wise organization observed in the main analysis is not unique to Qwen2.5. Across ChatGLM3, Mistral, OLMo, and Qwen2.5, relatively few SER neurons are selected in the earliest layers, followed by a marked increase toward intermediate depth and a broader distribution across later layers. At the same time, layers containing more SER neurons do not necessarily have larger mean standardized DiD scores. High selectivity can occur even in layers containing relatively few selected neurons. We therefore treat *SER density* and *SER selectivity* as complementary properties of the response-side representation rather than as interchangeable measures.

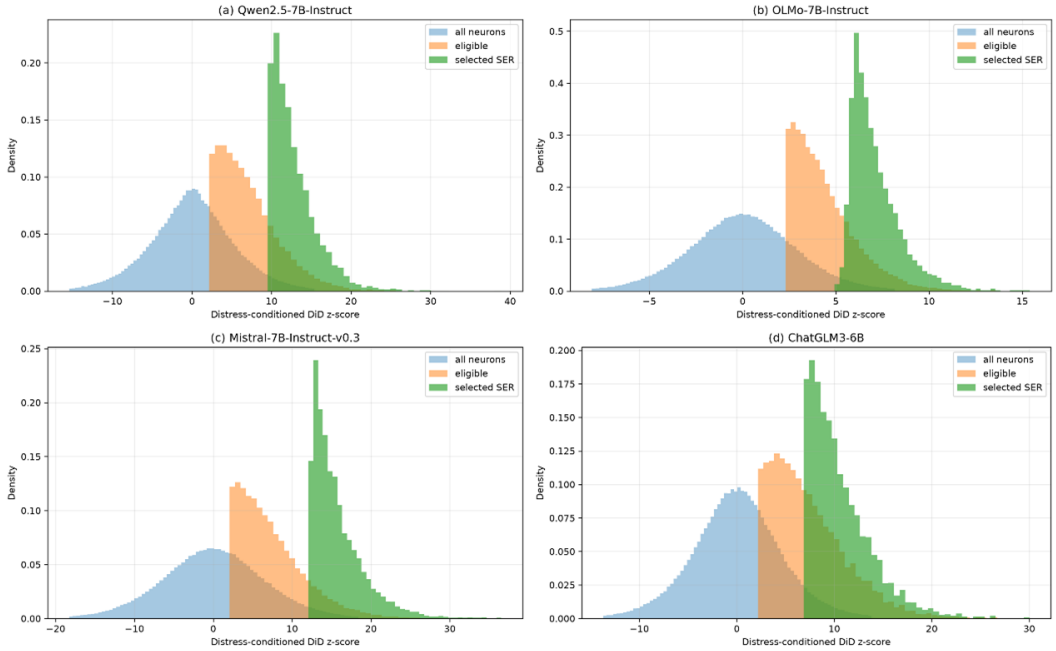


Figure 4: **Discovery-stage distributions of distress-conditioned standardized DiD statistics across model families.** Panels show Qwen2.5-7B-Instruct, OLMo-7B-Instruct, Mistral-7B-Instruct-v0.3, and ChatGLM3-6B. Blue distributions contain all FFN neurons, orange distributions contain neurons satisfying the pre-specified eligibility criteria, and green distributions contain the final selected SER neurons. In every model, selection shifts the population toward strongly positive distress-conditioned effects. Because the plotted distributions arise from the same discovery data used for selection, this figure is treated as a descriptive diagnostic rather than independent evidence of generalization.

C Functional Characterization Details

C.1 Functional-Characterization Samples and Extraction

ESConv sample construction. For the affective–instrumental analysis, we retain 8,413 supporter turns: 2,786 Low+Affective, 2,894 Low+Instrumental, 1,465 High+Affective, and 1,268 High+Instrumental. The data are split by conversation into 5,932 test and 2,481 calibration turns. The distress threshold is the calibration-set median of initial emotion intensity (4.0) and is re-estimated outside each held-out fold during cross-fitting. Turns labeled Question, Self-disclosure, Restatement, or Other (3,801, 1,713, 1,089, and 3,341 turns, respectively) are excluded from activation extraction because they do not belong to the pre-specified affective or instrumental strategy groups.

EPITOME sample construction. For functional characterization, EPITOME contributes 2,988 examples with complete annotations for Emotional Reaction, Interpretation, and Exploration, comprising 2,088 test and 900 calibration examples. This subset is distinct from the 3,023 EPITOME examples used for the causal intervention analyses in Section D.3.

Activation extraction. For both ESConv and EPITOME, extraction processes the prompt, truncated to at most 768 tokens, together with the first eight response tokens in a single forward pass with `use_cache=False`. For each neuron, recruitment is computed as

$$\Delta A_i = \text{mean}\left(h_{1:8,i}^{\text{response}}\right) - h_i^{\text{prompt-end}}, \tag{14}$$

using all available response states when fewer than eight response tokens are present. The resulting sample-level values are subsequently normalized out-of-fold before aggregation over the frozen neuron set.

C.2 Cross-Fitted Normalization and Statistical Uncertainty

ESConv samples are grouped by conversation ID, and EPITOME samples are grouped by source ID. We use five grouped folds and repeat partitioning with random seeds 42, 3407, and 2027.

For each held-out fold, normalization statistics

$$\mu_i^{(-k)}, \qquad \sigma_i^{(-k)} \tag{15}$$

are estimated only from samples outside that fold.

For ESConv, the high/low distress threshold is similarly estimated outside the held-out fold:

$$\tau^{(-k)} = \text{Median}\left\{d_c : c \notin k\right\}, \tag{16}$$

where d_c is the initial emotion intensity of conversation c.

Confidence intervals are obtained with conversation-cluster bootstrap for ESConv and source-cluster bootstrap for EPITOME, using 1,000 repetitions.

C.3 Mechanism-Level Interpretation

Emotional Reaction is emphasized because it is the EPITOME mechanism most closely aligned with affective empathic response realization. Interpretation provides a weaker complementary association, whereas Exploration is weak or negative. This mechanism-level heterogeneity is important: SER should not be interpreted as an undifferentiated scalar representation of all forms of empathy.

C.4 Factorial Response Construction Protocol

Contexts. We sample 1,000 unique ESConv conversation contexts (`selected_contexts.jsonl`, seed 42), each providing the full conversation history and the last user utterance as the shared prompt for all four conditions. Sampling is deduplicated by conversation so that no conversation contributes more than one context.

Constructor model and decoding. All four responses are generated by a single constructor model, Qwen2.5-7B-Instruct, in bfloat16 with seed 42 and greedy decoding (`do_sample=False`, at most 512 new tokens, input truncated to 2,048 tokens). The four evaluated models never generate their own responses: the same fixed A/B/C/D quadruplets are scored by every model family under teacher forcing, so cross-model differences reflect frozen SER recruitment rather than generation differences. JSON parsing accepts either a raw object or an object embedded in markdown fences; up to three generation attempts are allowed per context, and all 1,000 requested quadruplets were produced with zero generation failures.

Factor definitions. The 2×2 design crosses empathic acknowledgment with instrumental support: Empathy+Support (A: empathy 1, support 1), Empathy-only (B: 1/0), Support-only (C: 0/1), and Neither (D: 0/0).

Generation prompt. The constructor receives the following system instruction and user template verbatim (with {`context`} replaced by the conversation):

```
[System]
You construct matched assistant responses for a controlled
↪ neuroscience
experiment.
Return only valid JSON with keys A, B, C, D. Do not include markdown.
The four responses must preserve the user's facts and topic, be polite
↪ and
coherent, and be minimally matched in sentence count and length.
Each response must be exactly one sentence and should be 18-26 words
↪ if
English, or similarly concise if another language.
A must contain both empathy/affective acknowledgment and
```

```
support/reassurance/guidance.
B must contain empathy/affective acknowledgment only, with no
↪ reassurance,
encouragement, advice, solution, or action plan.
C must contain support/reassurance/guidance only, with no explicit
↪ emotion
acknowledgment, reflection of feelings, or validation.
D must be relevant and factual only, with no empathy, validation,
reassurance, encouragement, advice, action plan, or emotional support.

[User]
Conversation context:
{context}

Create four matched assistant responses.
Use the same language as the context if it is clear; otherwise use
↪ English.
Keep each response exactly one sentence. Keep all four lengths close;
↪ do not
make A systematically longer than B, C, or D.
Use the same concrete topic and referents in all four responses.
Return exactly:
{"A": "...", "B": "...", "C": "...", "D": "..."}
```

Validation protocol. Every quadruplet undergoes rule-based validation (`validate_quadruplets.py`) along three axes: (i) factor presence, checked with bilingual (English/Chinese) keyword lists for empathy markers (e.g. `feel`, `understand`, `sounds like`, `it makes sense`) and support markers (e.g. `you can`, `try`, `plan`, `reach out`) against the expected $(1,1)/(1,0)/(0,1)/(0,0)$ pattern; (ii) length matching, flagging quadruplets whose max/min token-count ratio exceeds 1.50; and (iii) structural matching, flagging sentence-count spreads larger than one and recording all six pairwise lexical Jaccard similarities. The keyword checks are deliberately conservative diagnostics: they do not delete any sample, and all $1{,}000$ quadruplets are retained for analysis. A random 100-context subset with full A/B/C/D texts (`manual_check_100.csv`, seed 42) is exported for human inspection. Early-response SER recruitment is then measured with the same first-8-token readout and cross-fitting procedure as in Sections C.1 and C.2.

C.5 Additional Analyses: Disentangling Empathy and Support

We provide additional analyses for the factorial response-control experiment introduced in Section 5.3. Figure 6 reports the complete condition-wise recruitment profiles, while Figure 7 examines whether the same pattern persists across model depth.

Full condition-wise recruitment. Figure 6 reports the complete frozen-SER recruitment profiles for all four factorial conditions.

Layer-wise robustness. To determine whether the factorial pattern is driven by a narrow depth range, we partition frozen SER neurons into lower, middle, and upper layer groups and recompute condition-wise early-response recruitment within each group.

D Causal Intervention Evaluation and Robustness

D.1 EPITOME Intervention Scoring

We follow the released EPITOME implementation of Sharma et al. (2020). Specifically, we use the `roberta-base` tokenizer/backbone with the repository's `BiEncoderAttentionWithRationaleClassification` architecture and three separately trained Reddit mechanism classifiers. The checkpoints correspond to Emotional Reaction, Interpretation, and Exploration and are saved by the released training recipe as

`output/reddit_ER.pth`, `output/reddit_IP.pth`, and `output/reddit_EX.pth`, respectively. The seeker post and response are tokenized separately, padded or truncated to 64 tokens, and passed through the two encoder branches. We use the empathy-identification logits from the mechanism-specific classification head; the rationale-extraction logits are not used in our intervention endpoint.

For each EPITOME mechanism

$$m \in \{\mathrm{ER}, \mathrm{IP}, \mathrm{EX}\}, \tag{17}$$

the corresponding classifier produces logits z_m over annotation levels $c \in \{0, 1, 2\}$. We convert them to class probabilities

$$p_m(c) = \mathrm{softmax}(z_m)_c \tag{18}$$

and compute the expected mechanism score

$$E_m = \sum_{c=0}^{2} c\, p_m(c). \tag{19}$$

The EPITOME score reported in the intervention analyses is

$$Y_{\mathrm{EPITOME}} = \frac{E_{\mathrm{ER}} + E_{\mathrm{IP}} + E_{\mathrm{EX}}}{3}. \tag{20}$$

This intervention endpoint is distinct from the functional characterization in Section C, where Emotional Reaction is examined separately because it provides the closest mechanism-level correspondence to SER.

D.2 Other Evaluation Benchmarks

EmoBench, ARC-Challenge, and MMLU use exact multiple-choice accuracy. EQ-Bench follows its released normalized scoring procedure.

BLEU and PPL are retained as generation-level controls. For PPL, lower values indicate better language-modeling performance, so a positive intervention delta corresponds to degradation.

D.3 Matched-Neuron Causal Controls on EPITOME

To examine whether the intervention pattern is specific to the identified response-side population, we apply the same activation-scaling intervention to the frozen layer-matched random, high-activation, and distress-control neuron sets. Each control set has the same cardinality and layer histogram as the corresponding frozen Top-1% SER set. All conditions are evaluated on the same 3,023 EPITOME examples. For each model and neuron set, $\alpha = 1$ denotes the unmodified model, $\alpha < 1$ suppresses the selected neurons, and $\alpha > 1$ amplifies them.

SER exhibits a consistent bidirectional intervention pattern. Across all four models, SER suppression decreases EPITOME and amplification increases it, including at $\alpha = 0.5$ and $\alpha = 1.5$. Layer-matched random sets produce smaller and less consistent changes, so matched cardinality and depth alone do not reproduce the SER pattern.

Suppression magnitude alone is not specific to SER. Complete suppression of high-activation neurons can reduce EPITOME more than SER suppression for OLMo, Mistral, and ChatGLM3. However, high-activation and distress-control sets do not reproduce the same cross-model bidirectional pattern; their amplification effects vary in sign and magnitude. The causal claim therefore rests on matched controls, functional specificity, cross-model consistency, and bidirectional intervention rather than suppression magnitude alone.

D.4 SER Causal Effects on ESConv Strategy Margin

To complement the EPITOME intervention analysis, we examine whether manipulating frozen SER neurons also changes the type of support expressed during generation. We define the ESConv strategy margin as the summed probability of affective support strategies minus that of instrumental

Table 6: **EPITOME scores under interventions on SER and matched control neuron sets.** Each cell reports the absolute EPITOME score, with the change relative to the corresponding unmodified model ($\alpha = 1$) shown in parentheses. All conditions use frozen Top-1% neuron sets and the same 3,023 EPITOME examples.

Model	Neuron set	$\alpha = 0$	$\alpha = 0.5$	$\alpha = 1$	$\alpha = 1.5$	$\alpha = 2$
Qwen2.5-7B-Instruct	SER	.4620 (−.2161)	.5574 (−.1207)	.6781	.7129 (+.0348)	.7205 (+.0424)
	Layer-matched random	.6773 (−.0008)	.6776 (−.0005)	.6781	.6720 (−.0061)	.6686 (−.0095)
	High-activation	.6255 (−.0526)	.6067 (−.0714)	.6781	.6474 (−.0307)	.6267 (−.0514)
	Distress-control	.6066 (−.0715)	.6090 (−.0691)	.6781	.6825 (+.0044)	.6660 (−.0121)
OLMo-7B-Instruct	SER	.6401 (−.0745)	.6957 (−.0189)	.7145	.7183 (+.0037)	.7259 (+.0114)
	Layer-matched random	.7070 (−.0075)	.7101 (−.0045)	.7145	.7159 (+.0014)	.7215 (+.0070)
	High-activation	.2811 (−.4335)	.6822 (−.0323)	.7145	.7492 (+.0347)	.7600 (+.0455)
	Distress-control	.5847 (−.1299)	.6966 (−.0180)	.7145	.7182 (+.0036)	.7166 (+.0020)
Mistral-7B-Instruct-v0.3	SER	.4711 (−.1678)	.5725 (−.0664)	.6389	.6545 (+.0156)	.6621 (+.0231)
	Layer-matched random	.6372 (−.0017)	.6403 (+.0013)	.6389	.6408 (+.0018)	.6425 (+.0035)
	High-activation	.4037 (−.2352)	.6148 (−.0242)	.6389	.6390 (+.0001)	.6300 (−.0089)
	Distress-control	.6274 (−.0115)	.6280 (−.0110)	.6389	.6377 (−.0013)	.6237 (−.0153)
ChatGLM3-6B	SER	.4935 (−.1092)	.5550 (−.0478)	.6027	.6339 (+.0312)	.6590 (+.0562)
	Layer-matched random	.6049 (+.0022)	.6050 (+.0022)	.6027	.6061 (+.0033)	.6085 (+.0058)
	High-activation	.3953 (−.2074)	.5762 (−.0265)	.6027	.6002 (−.0025)	.5786 (−.0242)
	Distress-control	.5956 (−.0071)	.6013 (−.0014)	.6027	.6123 (+.0096)	.6201 (+.0174)

Table 7: **Paired SER intervention effects on the ESConv strategy margin.** Affective score is the summed probability of *Reflection of Feelings* and *Affirmation/Reassurance*; instrumental score is the summed probability of *Providing Suggestions* and *Information*. Margin is defined as affective minus instrumental score. Values are paired changes from the corresponding unmodified model ($\alpha = 1$) over 2,518 ESConv intervention contexts. Bracketed values give 95% paired bootstrap confidence intervals for the margin change.

Model	$\alpha = 0$ (suppression)			$\alpha = 2$ (amplification)		
	Aff.	Inst.	Margin	Aff.	Inst.	Margin
Qwen2.5-7B	−.049	+.050	−.099 [−.114, −.085]	+.058	−.050	+.107 [+.093, +.122]
OLMo-7B	−.097	+.099	−.196 [−.215, −.177]	+.040	−.049	+.088 [+.071, +.107]
Mistral-7B	−.058	+.041	−.099 [−.117, −.083]	+.088	−.068	+.156 [+.139, +.175]
ChatGLM3-6B	−.021	+.016	−.037 [−.052, −.022]	+.016	−.046	+.062 [+.048, +.076]

support strategies. Specifically, the affective score combines *Reflection of Feelings* and *Affirmation/Reassurance*, whereas the instrumental score combines *Providing Suggestions* and *Information*. Table 7 reports paired intervention effects on this margin under complete SER suppression ($\alpha = 0$) and amplification ($\alpha = 2$). All four model families are evaluated on the same 2,518 ESConv intervention contexts. Confidence intervals are obtained with 1,000 paired bootstrap repetitions over contexts.

Across all four model families, complete SER suppression produces a negative strategy-margin shift, indicating reduced affective support relative to instrumental support. Conversely, amplification produces a positive margin shift in every model. All corresponding 95% bootstrap confidence intervals exclude zero. This bidirectional strategy-level effect mirrors the EPITOME intervention pattern and provides complementary evidence that SER activity causally influences how empathic support is realized during generation.

D.5 Mechanism-Level Effects of Depth-Restricted SER Suppression

To examine whether the causal effect of SER is distributed uniformly across empathy mechanisms and model depth, we suppress either the complete frozen SER set or depth-restricted subsets corresponding to the bottom, middle, and top portions of the network. This analysis is conducted on Qwen2.5-7B-Instruct using the same 3,023 EPITOME examples as the main intervention analysis.

Table 8: **Mechanism-level effects of full and depth-restricted SER suppression on EPITOME.** Values report the paired change relative to the unmodified model. Negative values indicate a reduction in the corresponding EPITOME mechanism score. Confidence intervals are 95% intervals over the 3,023 evaluation examples.

Suppressed set	EPITOME mechanism	Effect	95% CI	p
All SER	Emotional Reaction	$-.5916$	$[-.6032, -.5790]$	$< .001$
	Interpretation	$-.0860$	$[-.1006, -.0723]$	$< .001$
	Exploration	$-.0088$	$[-.0215, +.0041]$	.156
Bottom	Emotional Reaction	$-.0607$	$[-.0698, -.0515]$	$< .001$
	Interpretation	$-.0305$	$[-.0431, -.0178]$	$< .001$
	Exploration	$+.0457$	$[+.0338, +.0594]$	$< .001$
Middle	Emotional Reaction	$-.1207$	$[-.1310, -.1105]$	$< .001$
	Interpretation	$-.0414$	$[-.0531, -.0292]$	$< .001$
	Exploration	$-.0113$	$[-.0211, -.0014]$	.030
Top	Emotional Reaction	$-.3648$	$[-.3761, -.3530]$	$< .001$
	Interpretation	$+.0474$	$[+.0320, +.0633]$	$< .001$
	Exploration	$-.0328$	$[-.0426, -.0230]$	$< .001$

We report paired changes relative to the unmodified model separately for Emotional Reaction, Interpretation, and Exploration.

The intervention effect is concentrated on Emotional Reaction. As shown in Figure 8 and Table 8, suppressing the full SER set strongly reduces Emotional Reaction ($\Delta = -.5916$), with much smaller changes in Interpretation and Exploration. The same asymmetry appears under depth-restricted suppression: the largest observed Emotional Reaction reduction occurs for the top-layer subset ($-.3648$), followed by the middle ($-.1207$) and bottom ($-.0607$) subsets. Interpretation and Exploration exhibit smaller and less uniform effects; top-layer suppression increases Interpretation, whereas bottom-layer suppression increases Exploration. Thus, the causal effect of SER suppression is not uniform across EPITOME mechanisms, but is most strongly associated with Emotional Reaction.

D.6 Full Cross-Model Top-p% Ablation Results

To complement the compact summary in Table 4, we report the complete Top-p% ablation results for all four model families. For each masking fraction, we provide the number of masked FFN neurons, the absolute EPITOME score after masking, the paired change from the corresponding unablated baseline, and its 95% bootstrap confidence interval. The final column reports the change in ΔEPITOME relative to the immediately preceding masking fraction and is included only as a descriptive diagnostic of the non-monotonic dependence on p. All conditions contain $N = 3023$ evaluation examples. Because this extended ablation analysis uses a separately retrained RoBERTa evaluator, its absolute baseline scores are not directly comparable to those reported for the main causal-intervention experiment.

Cross-model pattern. All four model families show negative paired EPITOME changes at every tested masking fraction, with all reported 95% bootstrap confidence intervals excluding zero. The effect is already present at the top 0.1% of the ranking. Qwen shows a pronounced increase in effect magnitude around the Top-1% threshold and reaches its largest observed reduction at 3%. Mistral reaches a comparatively large reduction by 0.2% and remains broadly stable thereafter, whereas OLMo and ChatGLM3 exhibit smaller but persistent effects. The marginal changes alternate in sign at several masking fractions, confirming that the relationship between masking fraction and effect magnitude is not strictly monotonic.

Table 9: **Detailed Top-$p\%$ ablation results for Qwen2.5-7B.** The unablated EPITOME baseline is .6777. Δ denotes the paired change from this baseline, and Δ_{marg} denotes the change relative to the preceding masking fraction.

Top-$p\%$	# Neurons	Masked	ΔEPITOME	95% CI	Δ_{marg}
0.1	531	.5378	-.1398	$[-.1463, -.1336]$	–
0.2	1,061	.5385	-.1391	$[-.1453, -.1329]$	+.0007
0.3	1,592	.5364	-.1413	$[-.1477, -.1355]$	-.0022
0.4	2,122	.5544	-.1233	$[-.1290, -.1171]$	+.0180
0.5	2,653	.5327	-.1450	$[-.1511, -.1385]$	-.0217
0.6	3,183	.5284	-.1493	$[-.1552, -.1434]$	-.0043
0.7	3,714	.5346	-.1430	$[-.1495, -.1365]$	+.0062
0.8	4,244	.5247	-.1530	$[-.1587, -.1461]$	-.0099
0.9	4,774	.5185	-.1592	$[-.1656, -.1530]$	-.0062
1.0	5,305	.4640	-.2137	$[-.2208, -.2073]$	-.0545
2.0	10,609	.4465	-.2311	$[-.2382, -.2243]$	-.0174
3.0	15,913	.4230	-.2546	$[-.2613, -.2480]$	-.0235
4.0	21,218	.4298	-.2478	$[-.2545, -.2407]$	+.0068
5.0	26,522	.4389	-.2388	$[-.2458, -.2319]$	+.0090

Table 10: **Detailed Top-$p\%$ ablation results for Mistral-7B.** The unablated EPITOME baseline is .6652.

Top-$p\%$	# Neurons	Masked	ΔEPITOME	95% CI	Δ_{marg}
0.1	459	.5246	-.1405	$[-.1468, -.1342]$	–
0.2	918	.4543	-.2109	$[-.2160, -.2054]$	-.0704
0.3	1,377	.4583	-.2069	$[-.2123, -.2016]$	+.0040
0.4	1,836	.4497	-.2155	$[-.2208, -.2103]$	-.0086
0.5	2,294	.4442	-.2209	$[-.2264, -.2153]$	-.0054
0.6	2,753	.4558	-.2093	$[-.2145, -.2037]$	+.0116
0.7	3,212	.4495	-.2156	$[-.2209, -.2103]$	-.0063
0.8	3,671	.4501	-.2150	$[-.2201, -.2096]$	+.0006
0.9	4,129	.4593	-.2058	$[-.2119, -.1998]$	+.0092
1.0	4,588	.4521	-.2131	$[-.2185, -.2079]$	-.0073
2.0	9,176	.4749	-.1902	$[-.1965, -.1847]$	+.0229
3.0	13,763	.4724	-.1927	$[-.1981, -.1864]$	-.0025
4.0	18,351	.4540	-.2111	$[-.2169, -.2053]$	-.0184
5.0	22,938	.4483	-.2169	$[-.2235, -.2109]$	-.0058

E Reproducibility Configuration

Each architecture is analyzed independently. Worker-specific random streams are deterministic functions of the base seed and worker rank, and each worker holds a complete model copy.

The experiment outputs retain the model configuration, frozen neuron identities, sample-level activation summaries, paired intervention outputs, and aggregate evaluation statistics required to reconstruct the reported results.

Table 11: **Detailed Top-$p\%$ ablation results for OLMo-7B.** The unablated EPITOME baseline is .8174.

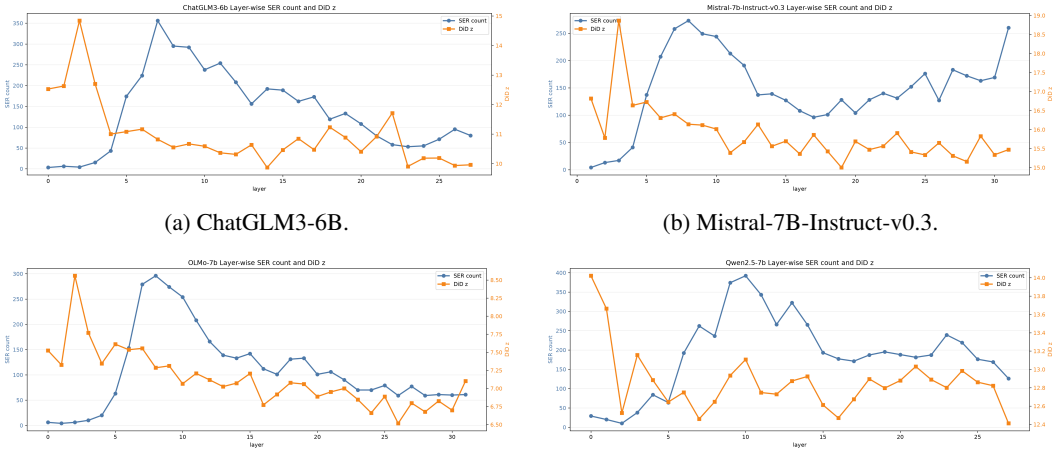
Top-$p\%$	# Neurons	Masked	ΔEPITOME	95% CI	Δ_{marg}
0.1	353	.7378	-.0795	$[-.0865, -.0724]$	–
0.2	705	.7482	-.0691	$[-.0761, -.0624]$	+.0104
0.3	1,057	.7478	-.0696	$[-.0773, -.0622]$	-.0005
0.4	1,410	.7477	-.0696	$[-.0776, -.0621]$	-.0001
0.5	1,762	.7342	-.0831	$[-.0911, -.0749]$	-.0135
0.6	2,114	.7434	-.0740	$[-.0819, -.0661]$	+.0091
0.7	2,466	.7214	-.0959	$[-.1039, -.0880]$	-.0220
0.8	2,819	.7046	-.1127	$[-.1211, -.1043]$	-.0168
0.9	3,171	.7135	-.1039	$[-.1116, -.0957]$	+.0089
1.0	3,523	.7102	-.1072	$[-.1152, -.0991]$	-.0033
2.0	7,046	.7343	-.0830	$[-.0907, -.0755]$	+.0241
3.0	10,568	.7120	-.1054	$[-.1135, -.0969]$	-.0224
4.0	14,091	.7088	-.1086	$[-.1167, -.1005]$	-.0032
5.0	17,613	.7248	-.0926	$[-.1005, -.0850]$	+.0160

Table 12: **Detailed Top-$p\%$ ablation results for ChatGLM3-6B.** The unablated EPITOME baseline is .6555.

Top-$p\%$	# Neurons	Masked	ΔEPITOME	95% CI	Δ_{marg}
0.1	384	.5971	-.0584	$[-.0652, -.0512]$	–
0.2	767	.5973	-.0582	$[-.0653, -.0507]$	+.0002
0.3	1,151	.6048	-.0507	$[-.0576, -.0438]$	+.0075
0.4	1,534	.6015	-.0540	$[-.0611, -.0465]$	-.0033
0.5	1,918	.5984	-.0570	$[-.0643, -.0498]$	-.0031
0.6	2,301	.5748	-.0806	$[-.0884, -.0727]$	-.0236
0.7	2,685	.5710	-.0844	$[-.0921, -.0777]$	-.0038
0.8	3,068	.5670	-.0885	$[-.0961, -.0810]$	-.0040
0.9	3,452	.5574	-.0980	$[-.1055, -.0907]$	-.0096
1.0	3,835	.5660	-.0894	$[-.0967, -.0812]$	+.0086
2.0	7,670	.5562	-.0993	$[-.1072, -.0910]$	-.0098
3.0	11,505	.5609	-.0946	$[-.1022, -.0869]$	+.0047
4.0	15,340	.5511	-.1044	$[-.1123, -.0967]$	-.0098
5.0	19,175	.5576	-.0978	$[-.1059, -.0897]$	+.0065

Table 13: **Core experiment configuration.**

Setting	Value
Base seed	42
Source split	70% discovery / 30% held-out
Discovery folds	5
Functional-analysis folds	5
Functional-analysis seeds	42, 3407, 2027
Frozen SER cardinality	1% of all FFN neurons, selected from the eligible pool
Counterfactual total-token budget	672
Functional prompt budget	768
Intervention input budget	1,024
Generative max new tokens	256
Short-answer max new tokens	16
Bootstrap repetitions	1,000
Inference dtype	bfloat16
Hardware	4 × NVIDIA A800



(a) ChatGLM3-6B.

(b) Mistral-7B-Instruct-v0.3.

(c) OLMo-7B-Instruct.

(d) Qwen2.5-7B-Instruct.

Figure 5: **Layer-wise distribution and selectivity of frozen SER neurons across model families.** Blue curves show the number of selected SER neurons in each layer, while orange curves show their mean standardized DiD score. Across architectures, selected neurons are sparse in the earliest layers, become substantially more numerous at intermediate depths, and remain distributed through later layers. The layer-wise profile of SER count does not coincide with that of mean standardized DiD, indicating that representation density and response selectivity capture distinct properties of the identified SER population.

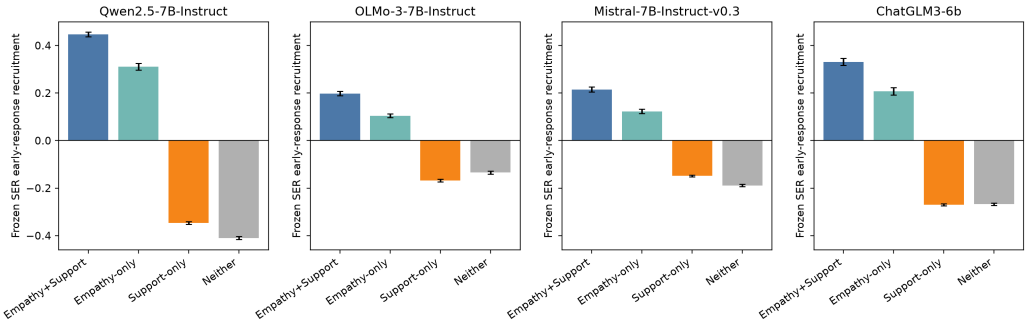


Figure 6: **Full condition-wise frozen SER recruitment.** For each of 1,000 ESConv conversation contexts, we construct matched responses that independently manipulate empathic acknowledgment and support: EMPATHY+SUPPORT (A), EMPATHY-ONLY (B), SUPPORT-ONLY (C), and NEITHER (D). Across all four model families, EMPATHY+SUPPORT yields the highest early-response SER recruitment, followed by EMPATHY-ONLY, whereas responses without empathic acknowledgment show substantially lower recruitment. Error bars denote 95% conversation-cluster bootstrap confidence intervals over 1,000 unique contexts. Because recruitment scores are normalized relative to the prompt-end baseline, their absolute signs are not interpreted as excitation or inhibition; our conclusions rely on matched within-context contrasts.

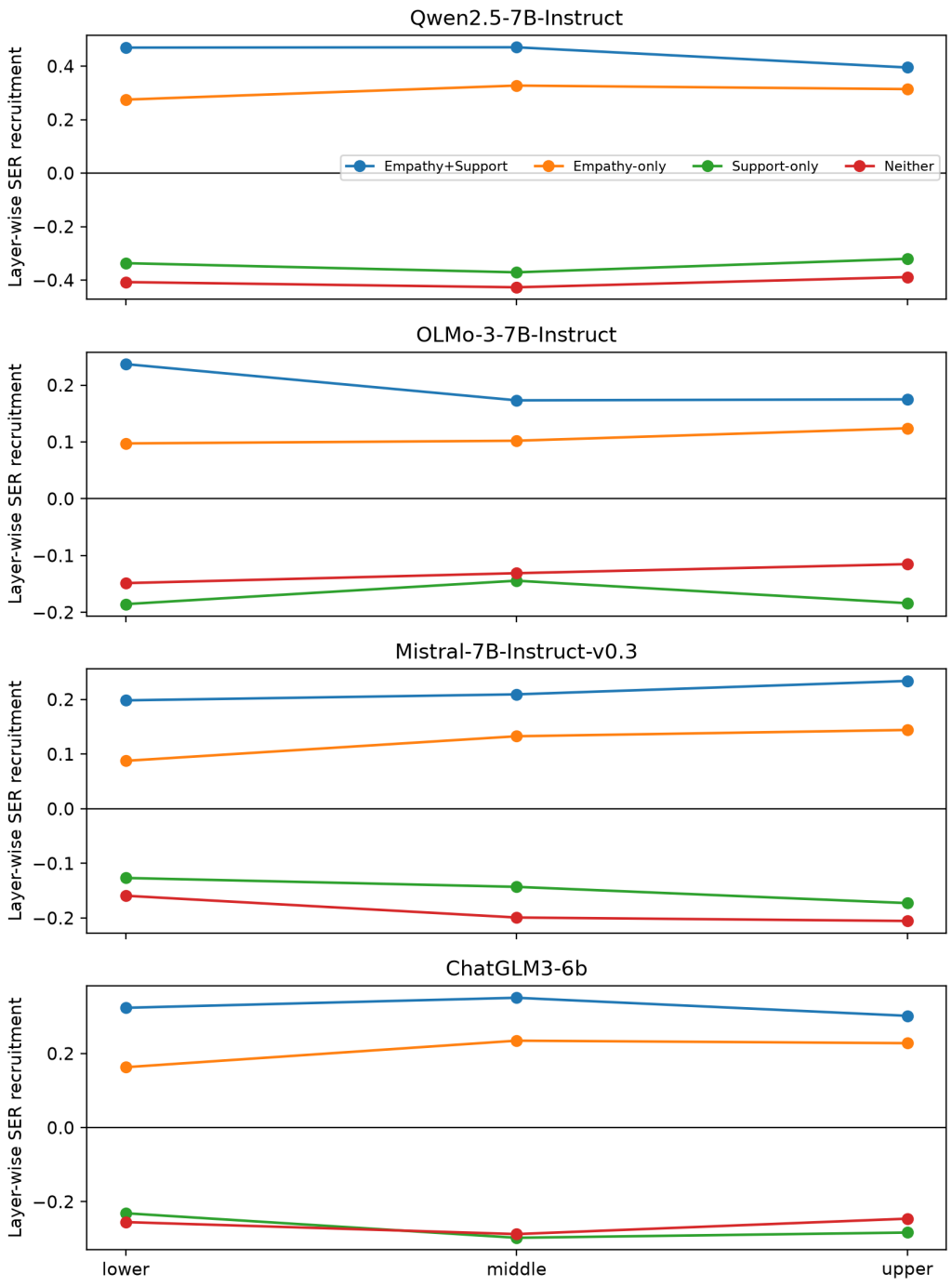


Figure 7: **Layer-wise robustness of the empathy–support factorial pattern.** Frozen SER neurons are partitioned into lower, middle, and upper depth groups, and condition-wise early-response recruitment is recomputed separately within each group. The qualitative separation between empathy-containing responses (EMPATHY+SUPPORT and EMPATHY-ONLY) and responses without empathic acknowledgment (SUPPORT-ONLY and NEITHER) persists across depth groups and model families. Moreover, EMPATHY+SUPPORT generally recruits SER more strongly than EMPATHY-ONLY, indicating that the additional contribution of support is not confined to a single depth region. Error bars denote 95% conversation-cluster bootstrap confidence intervals over 1,000 unique contexts.

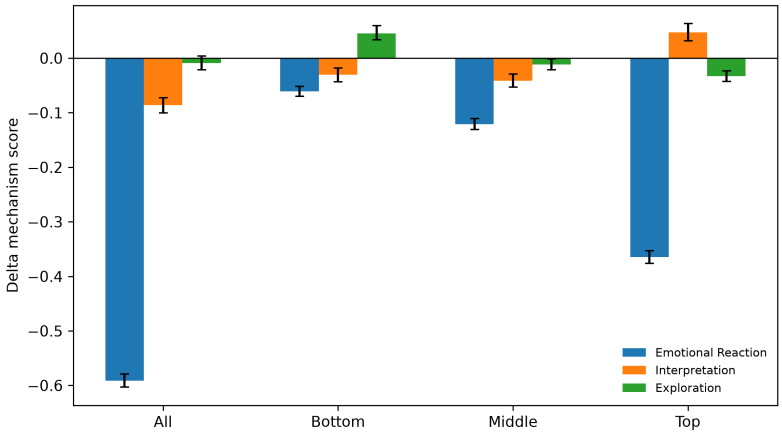


Figure 8: **Mechanism-level effects of depth-restricted SER suppression on Qwen2.5-7B-Instruct.** Frozen SER neurons are partitioned into bottom-, middle-, and top-layer subsets, and each subset is suppressed separately. Bars report paired changes relative to the unmodified model for the three EPITOME mechanisms: Emotional Reaction, Interpretation, and Exploration. Suppression produces its largest and most consistent reduction in Emotional Reaction, with the strongest observed effect arising from the top-layer subset. Interpretation and Exploration show smaller and more heterogeneous changes across depth groups. Error bars denote 95% confidence intervals over the same 3,023 EPITOME examples used in the intervention analysis.