Emotion Is Not External Knowledge: Internal Attention Signals for Affective State Reasoning

Anonymous ACL submission

Abstract

Large Language Models (LLMs) increasingly operate in multi-turn interactions that require maintaining latent conversational states, not only explicit facts. We study affective state as an implicit state inferred from sparse contextual cues that remain relevant after topic shifts, and ask whether LLMs expose internal attention signals for routing such evidence at inference time. Prior attention interventions rely on labeled in-domain head profiles and externally specified target spans, limiting transfer across affective-evidence distributions. We propose Self-targeted Internal Attention Boosting (SIAB), which localizes sparse affective heads via causal localization (in detail activation patching) on synthetic minimal pairs and boosts the content tokens these heads already attend to; both edited heads and target tokens are derived from the model's own behavior, enabling the same head set to be reused across tasks.We evaluate on ECC and GoEmotions across three LLMs and multiple prompting settings. Across settings, SIAB improves prompt-matched baselines by up to +10.3 macro-F1 on ECC and +5.5 on GoEmotions. Combined with PASTA-style external supervision, it achieves up to +13.84 accuracy points and +12.51 macro-F1 points, while mechanistic analyses trace the gains to sparse, emotion-enriched heads and show that layer proximity as a key factor behind additivity.

1 Introduction

Conversation is inherently stateful: What a user means at any point depends on what came before. As Large Language Models (LLMs) are increasingly deployed in such multi-turn interaction, successful responses must not only retain explicit factual context but also maintain latent conversational states, such as a user's frustration, uncertainty, or trust (Fazzi et al., 2025, Følstad et al., 2024). Affective state is a challenging case. It is rarely given as an explicit variable; instead, it must be inferred

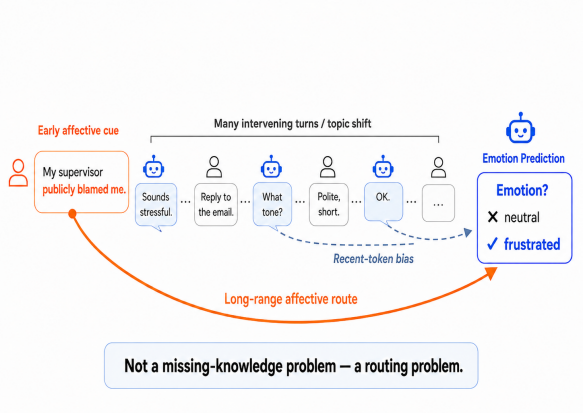

Figure 1: Long dialogues can bury early affective cues, turning affective prediction into a routing problem.

from sparse contextual cues whose relevance can persist across topic shifts. As illustrated in Figure 1, a model may therefore need to use an earlier affective cue when making a later prediction.

Unlike knowledge-intensive failures, where retrieval or tool use can supply a missing fact, affective continuity does not require adding new information. The relevant evidence is already in the dialogue; the difficulty is whether the model routes it to the final decision point. Emotion recognition therefore provides a controlled proxy for studying this routing problem at inference time.

Inference-time attention intervention is a natural way to probe this problem, since it reroutes existing contextual evidence during the forward pass, without parameter updates or newly supplied facts. Existing attention intervention methods, however, rely on external control signals. Post-hoc Attention STeering Approach (PASTA) (Zhang et al., 2024a) profiles which heads to edit from labeled in-domain data and steers attention toward externally specified spans. Later variants reduce but do not remove this reliance on external signals. These methods raise two questions: can internal attention signals route affective evidence without external spans **(RQ1)**?

and are internal and external signals redundant or complementary (**RQ2**)?

We address these questions with Self-targeted Internal Attention Boosting (SIAB), an attention intervention for affective state reasoning. It first localizes a sparse set of heads through causal localization via activation patching. It then uses these heads themselves to select target tokens. At inference time, SIAB adds a scalar bias to the pre-softmax attention logits of the localized heads, boosting the content tokens they already attend to. Both the edited heads and the target tokens thus derive from the model's own behavior.

We evaluate on two Emotion Recognition in Conversation (ERC) settings with complementary distributions of affective evidence. In the Emotion-Cause Conversation dataset (ECC; He et al., 2025), cues are dispersed across earlier turns and must survive later topic shifts; in GoEmotions (Demszky et al., 2020), a single dense cue must instead be disambiguated among many fine-grained categories. The same head set is reused across both.

Since our goal is not only to improve recognition but to understand the internal signal we exploit, we analyze when and why the localized heads work. The gains come from a sparse set of heads: a few heads are sufficient, and they attend more to affective words than random heads from early, middle, and late layers. Whether SIAB succeeds further depends on logit geometry and on whether the boosted tokens align with the final label decision. Finally, SIAB and PASTA compose according to layer-band proximity.

Our contributions are threefold: (1) we introduce SIAB, a self-targeted inference-time attention intervention that requires no parameter updates or annotated spans and composes with prompting and PASTA-style external supervision; (2) we show consistent gains across models, datasets, and prompting settings; and (3) we characterize its mechanism and boundary conditions.

2 Related Work

Existing attention-level methods externalize control to varying degrees, relying on dev-profiled heads with target tokens that are user-annotated (Zhang et al., 2024a), heuristically extracted (Zhang et al., 2024b, Yu et al., 2024), or applied at the head output rather than its attention (Li et al., 2023); residual-stream steering operates at a coarser granularity (Turner et al., 2025, Rimsky et al., 2024, Zou et al., 2023). These methods assume the relevant evidence can be externally identified, an assumption less natural for affective state, whose cues spread across turns, often lack lexical salience, and shift as the conversation unfolds. We instead derive both signals internally —heads from activation patching, target tokens from each head's own attention— and study how this internal signal composes with external ones.

Mechanistic interpretability provides tools for discovering internal structures that mediate model behavior. Activation patching (Wang et al., 2023, Goldowsky-Dill et al., 2023) has been used to localize concrete circuit components, such as name-mover and S-inhibition heads, through clean–corrupted activation contrasts, while direct logit attribution (Belrose et al., 2023) links intermediate contributions to final-token logits. These tools fit our goal of identifying internal signals for affective-evidence routing: rather than specifying emotion-bearing spans externally, we localize the attention heads that internally route affective contrast.

Emotion recognition in conversation naturally probes affective state across dialogue turns, yet existing mechanistic work operates at a coarser granularity: it studies sentiment directions (Tigges et al., 2024), mid-layer Multi-Head Self-Attention units (Tak et al., 2025), or residual-stream circuits for emotion *generation* (Wang et al., 2025), leaving head-level recognition unaddressed. Other work improves accuracy through retrieval, speaker modeling, or curriculum learning (Lei et al., 2023, Xue et al., 2024, Fu et al., 2025, Li et al., 2026), but treats the model as a black box and ignores its internal attention signal.

3 Method

We introduce SIAB, an inference-time attention intervention with two components: head-set construction via Activation Patching and heads refinement (§3.1, §3.2) and self-targeted pre-softmax logit boosting (§3.3). Figure 2 gives an overview. Both components derive their control signal from the model's own internal behavior.

We frame emotion prediction as generative label scoring. Given an input context X and a candidate label set $\mathcal{Y}$, we score each label using its answer-position verbalizer logit, rather than training a task-specific classification head.

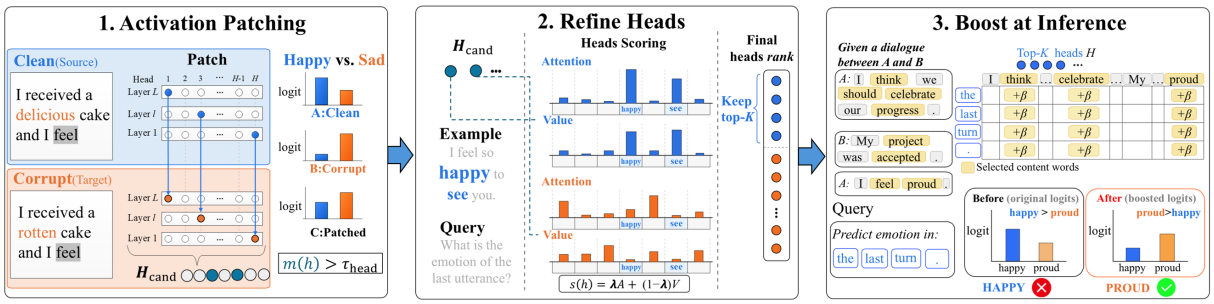

Figure 2: SIAB overview. Causal localization via activation patching selects sparse emotion-sensitive heads (4–20 of hundreds); SIAB adds a pre-softmax bias to the content tokens these heads already attend to at final query positions—without training, parameter updates, or external spans.

3.1 Head Selection via Activation Patching

Activation patching localizes the attention heads that causally affect a target prediction using prompt pairs that differ only in the feature driving that prediction. In our case, the difference in affective state is driven by the affective cues, so we build clean–corrupted pairs $(X^{\text{clean}}, X^{\text{corr}})$, e.g., X^{clean} = *"I received a delicious cake. I feel"* and X^{corr} = *"I received a rotten cake. I feel"* (Appendix A). Two properties make these pairs suitable for causal attribution. First, every prompt ends with `I feel`, so the model is induced to predict an emotion word at the final position, where we read off the effect. Second, the two prompts share the same syntactic structure and differ only in the affective cue, so any difference in the final-position emotion prediction can be attributed to that cue.

We run the model on the X^{corr} as a baseline. Then, for one head h at a time, we replace *only* that head's final-token output with the output it produced on the X^{clean} and complete the forward pass; everything else stays at its corrupted value. If h encodes the cue, restoring its clean output pulls the emotion prediction back toward the clean answer; if not, little changes. We quantify this as the normalized logit-difference recovery:

$$m(h) = \frac{\mathcal{L}(\text{patch}_h) - \mathcal{L}(X^{\text{corr}})}{\mathcal{L}(X^{\text{clean}}) - \mathcal{L}(X^{\text{corr}})}, \qquad (1)$$
$$\mathcal{L}(\cdot) = \text{logit}(e_{\text{clean}}) - \text{logit}(e_{\text{corr}}).$$

Here, e_{clean} and e_{corr} are the top-ranked final-position emotion words on X^{clean} and X^{corr}, respectively, and patch_h denotes the patched run described above. Thus $m(h) = 1$ when the swap fully restores the clean prediction and $m(h) = 0$ when it has no effect, following Belrose et al. (2023), Tigges et al. (2023). We apply the test to every head independently and retain those with $m(h) > \tau_{\text{head}}$; their union forms the candidate set $\mathcal{H}_{\text{cand}}$.

3.2 Label-free refinement on dialogues

Having obtained the candidate heads $\mathcal{H}_{\text{cand}}$, we do not directly treat all of them as suitable intervention targets. A head may obtain a high patching score $m(h)$ through generic mechanisms, such as positional tracking or induction-like copying (Olsson et al., 2022), rather than by routing affective cues. In natural dialogues, affective evidence is usually carried by lexical content. We therefore use a lightweight content-token filter that removes punctuation, special tokens, pronouns, interrogatives, and prepositions, while retaining negators and intensifiers . We then refine the candidates on dialogues by measuring their attention and value mass on content-word positions before the emotion prediction, using value mass because attention alone does not determine the information written to the residual stream (Kobayashi et al., 2020).

For this refinement stage, we use GPT-5.4 (OpenAI, 2026) to generate 200 dialogue samples (prompt in Appendix G). For each dialogue, we append a probe question, e.g., *"What is the emotion of the last utterance?"*, and inspect each candidate head immediately before the model generates the answer. No gold emotion labels are used.

For an arbitrary token position u, let $\alpha_{h,q,u}$ be the attention from query position $q \in \mathcal{Q}$ to u on head h, and let $v_{h,q,u}$ be the L2 norm of the corresponding value contribution. Because dialogue turns have different lengths, we normalize both quantities within the conversational turn as:

$$\tilde{\alpha}_{h,q,u} = \frac{\alpha_{h,q,u}}{\sum_{u' \in \text{turn}(u)} \alpha_{h,q,u'}}, \qquad (2)$$
$$\tilde{v}_{h,q,u} = \frac{v_{h,q,u}}{\sum_{u' \in \text{turn}(u)} v_{h,q,u'}},$$

where $\text{turn}(u)$ denotes all token positions in the

same conversational turn as u, including both content and non-content tokens. Thus, $\tilde{\alpha}_{h,q,u}$ and $\tilde{v}_{h,q,u}$ measure the fraction of a head's attention and value mass within that turn assigned to u.

Let $\mathcal{C}(r)$ denote the content-word positions in turn r. The candidate score is:

$$s_{\text{head}}(h) = \lambda A_h + (1-\lambda) V_h, \tag{3}$$

where A_h and V_h are computed by summing the turn-normalized masses over $u \in \mathcal{C}(r)$ and then averaging over turns r, query positions $q \in \mathcal{Q}$, and refinement dialogues. This yields the fraction of each head's turn-level mass assigned to substantive dialogue content.The final head set $\mathcal{H}$ is the top-K heads by s_{head}. Per-model values of K and the resulting head indices are reported in Appendix B.

3.3 Attention Logit Boosting

Crucially, the target tokens are not externally annotated, but selected from each chosen head's own attention distribution at the final query positions. At inference time, for each $(h, q) \in \mathcal{H} \times \mathcal{Q}$, let $\mathcal{T}_r(h, q)$ denote the top-r fraction of content-word tokens ranked by attention score $\alpha_{h,q,t}$. We add a scalar bias β to the pre-softmax attention logits $\ell_{h,q,t}$ of these tokens:

$$\ell'_{h,q,t} = \ell_{h,q,t} + \beta \mathbf{1}[t \in \mathcal{T}_r(h, q)]. \tag{4}$$

The attention distribution is then computed by applying softmax to the modified logits $\ell'_{h,q,t}$.

The head set $\mathcal{H}$ is fixed once and reused across both datasets without re-localization. Also the bias magnitude β is calibrated per model and then held fixed across both datasets. Hyperparameter values and the calibration protocol are in Appendix C.

4 Experimental Evaluation

4.1 Datasets

We evaluate on two benchmarks that stress different cue regimes. **ECC** (He et al., 2025) labels multi-turn dialogues (max. 831 tokens) with 10 fine-grained emotion classes, where affective evidence can appear early and must remain available across many turns. **GoEmotions** (Demszky et al., 2020) labels short Reddit comments (max. 52 tokens) with 27 multi-label emotion categories plus a Neutral class, where cues are local and information-dense but must be separated among many fine-grained labels. We exclude samples labeled *solely* as Neutral, since they contain no non-neutral affective cue for cue-targeting heads to act on; for multi-label samples, we retain the non-Neutral labels. This leaves 3,821 test examples for evaluation. Neutral results are reported in Appendix D.

4.2 Evaluation Metrics

We report macro-F1 and Accuracy. For ECC, we take the argmax over 10 labels. For GoEmotions, we follow the multi-label setting of Demszky et al. (2020); The temperature T and threshold t are tuned on the GoEmotions dev split (Appendix C). and top-1 accuracy counts the argmax label as correct if it belongs to the gold label set.

4.3 Models

We evaluate three open, small-to-mid-sized instruction-tuned LLMs from different model families: Gemma3-4B (Gemma Team, 2025), Qwen3-4B (Yang et al., 2025), and LLaMA3-8B (Grattafiori et al., 2024).

4.4 Baselines and Prompting Settings

We compare our method against three baselines.

The first is prompting: zero-shot and fixed few-shot prompting (templates in Appendix G).

The second is PASTA-style performance-profiled head selection, for which we report three variants. PASTA-ECC and PASTA-GoE are profiled on labeled ECC and GoEmotions dev-data respectively, probing in-domain effectiveness and cross-dataset transfer; PASTA-$\cap$ keeps only the heads shared by both profiles, testing whether a shared profile improves generality.We also report PASTA+SIAB to test complementarity with in-domain PASTA. PASTA's head count and variants are also tuned to their optimal values.

The third baseline, Layer-Matched Random, tests whether the effect is only layer-specific or due to within-layer redundancy: in either case, random same-layer heads should work as well (Voita et al., 2019, Michel et al., 2019). We sample 20 random subsets of non-selected heads from the same layers, matching our head-set size, and tune β on the same dev set.

4.5 Results

Table 1 reports results on two datasets. Δ are computed against the zero-shot baseline.

SIAB generalizes across models, datasets, and prompts. Across all 18 model–prompt–dataset settings, SIAB improves over the corresponding prompting baseline using the same head set, with

Table 1: Results on GoEmotions and ECC. *ZS*/*Mid*/*High* denote zero-/few-shot prompts (15/27 for GoEmotions; 5/10 for ECC). *PASTA-{ECC,GoE}* use heads selected on the named dataset; *PASTA-∩* keeps the shared top heads from these two profiles. *Rand.* is layer-matched random. Scores are $\times 100$; Δ is vs. ZS; best per model in **bold**.

		GoEmotions				ECC			
Model	**Config**	**Macro-F1**	**Top-1**	Δ_{F1}	Δ_{Top1}	**Acc.**	**Macro-F1**	Δ_{Acc}	Δ_{F1}
Gemma3	ZS	27.17	34.37	—	—	50.29	52.26	—	—
	ZS+SIAB	30.45	**38.15**	+3.28	+3.78	59.62	59.93	+9.32	+7.67
	Mid	30.55	33.45	+3.38	−0.92	56.48	55.51	+6.19	+3.25
	Mid+SIAB	32.65	36.68	+5.47	+2.30	59.90	60.15	+9.61	+7.89
	High	28.72	32.24	+1.55	−2.14	60.66	58.27	+10.37	+6.01
	High+SIAB	31.49	36.19	+4.32	+1.82	63.69	62.52	+13.40	+10.27
	PASTA-ECC	26.86	32.89	−0.31	−1.48	59.43	60.37	+9.14	+8.11
	PASTA-GoE	31.81	34.37	+4.64	+0.00	52.11	52.99	+1.82	+0.73
	PASTA-∩	29.19	34.65	+2.02	+0.28	59.12	60.02	+8.83	+7.76
	PASTA+SIAB	**33.75**	38.01	**+6.58**	+3.64	**64.13**	**64.77**	**+13.84**	**+12.51**
	Rand.	$26.01_{\pm 1.80}$	$32.13_{\pm 2.90}$	−1.16	−2.24	$41.89_{\pm 3.70}$	$45.60_{\pm 4.10}$	−8.40	−6.66
Qwen3	ZS	35.51	39.94	—	—	52.73	55.25	—	—
	ZS+SIAB	37.77	41.31	+2.26	+1.37	56.03	58.04	+3.30	+2.79
	Mid	35.56	40.38	+0.44	+0.44	55.63	58.02	+2.90	+2.78
	Mid+SIAB	**38.31**	42.69	**+2.79**	+2.75	57.71	59.91	+4.99	+4.66
	High	35.99	43.25	+0.48	+3.31	58.71	60.67	+5.98	+5.42
	High+SIAB	38.23	**44.26**	+2.71	**+4.32**	60.36	61.88	+7.63	+6.63
	PASTA-ECC	35.20	40.28	−0.31	+0.34	59.65	59.85	+6.92	+4.60
	PASTA-GoE	35.63	39.17	+0.12	−0.77	52.82	55.14	+0.09	−0.11
	PASTA-∩	35.99	39.80	+0.48	−0.14	57.65	59.11	+4.92	+3.86
	PASTA+SIAB	37.97	41.61	+2.46	+1.67	**62.76**	**63.09**	**+10.03**	**+7.84**
	Rand.	$34.64_{\pm 0.80}$	$39.75_{\pm 1.00}$	+0.03	−0.19	$52.39_{\pm 2.30}$	$55.27_{\pm 0.70}$	−0.34	+0.02
LLaMA3	ZS	30.39	33.24	—	—	59.27	60.28	—	—
	ZS+SIAB	33.64	**38.02**	+3.25	**+4.78**	61.80	63.23	+2.53	+2.95
	Mid	29.16	34.28	−1.23	+1.04	62.24	62.86	+2.96	+2.58
	Mid+SIAB	29.81	36.44	−0.58	+3.20	64.21	64.52	+4.94	+4.25
	High	26.57	32.53	−3.82	−0.72	64.97	64.81	+5.70	+4.53
	High+SIAB	27.55	35.85	−2.84	+2.61	**66.22**	**66.13**	**+6.95**	**+5.85**
	PASTA-ECC	23.20	26.31	−7.19	−6.93	65.28	63.35	+6.01	+3.07
	PASTA-GoE	32.14	31.51	+1.75	−1.73	62.09	61.74	+2.82	+1.46
	PASTA-∩	30.86	33.18	+0.47	−0.06	63.28	60.82	+4.01	+0.54
	PASTA+SIAB	**36.17**	37.73	**+5.78**	+4.49	65.48	63.55	+6.21	+3.27
	Rand.	$28.52_{\pm 2.80}$	$30.59_{\pm 3.20}$	−1.87	−2.65	$52.84_{\pm 5.00}$	$56.24_{\pm 4.10}$	−6.43	−4.04

gains up to $+10.27$ macro-F1 on ECC and $+5.47$ on GoEmotions. The effect is not limited to zero-shot: under 10-shot ECC, SIAB still adds $+1.21$ to $+4.25$ macro-F1 across models. On GoEmotions, it also reduces harmful few-shot effects: it turns Gemma3's Mid/High Top-1 drops into gains and weakens LLaMA3's Macro-F1 drops, while flipping its High-setting Top-1 change from negative to positive.

PASTA is strong in-domain but transfers poorly. In-domain PASTA profiles are competitive with, and often stronger than, ZS+SIAB. However, this strength is task-specific: applying PASTA-ECC heads to GoEmotions drops LLaMA3 by -7.19 macro-F1, and transfer in the reverse direction is similarly unstable. PASTA-∩ also loses in-domain performance without recovering cross-task gains. In contrast, SIAB uses the same causally selected head set across both datasets and gives positive Δ in all 18 settings. Thus, PASTA reaches a stronger in-domain peak when its profiling task matches the test task, whereas SIAB is weaker but more stable across tasks, motivating their composition.

The two signals compose additively in five of six (model, dataset) configurations. Across five configurations, PASTA+SIAB recovered 79–116% of the naive sum of the two individual zero-shot gains, and yielded the best macro-F1 on 4 of them, indicating the two signals operate largely orthogonally. The single exception, LLaMA3 on ECC, recovered only 54%. We attribute this to layer-band disjointness between the PASTA-

profiled and SIAB head sets and quantify the relationship in §5.4.

Effect of head selection: Evidence from Random. Layer-matched random head subsets matched or slightly trailed zero-shot on average. The ZS+SIAB vs. Random gap was at least $\times 1.6$ (max. $\times 2.6$) the Random std. on every cell. Since Random uses the same layers, head-set size, and tuning protocol, this argues against a simple accumulation of layer-level effects. It also shows that same-layer heads are not interchangeable for this intervention: the gain depends on the specific heads selected by SIAB.

5 Mechanistic Analysis

To analyze when and why the intervention succeeded, we partition test samples into four groups based on the joint behavior of Zero-Shot (ZS) and the proposed SIAB predictions. **POS**: ZS incorrect, SIAB correct, **NEG**: ZS correct, SIAB incorrect, **EASY**: both ZS and SIAB correct, **HARD**: both ZS and SIAB incorrect.

5.1 Sparsity of the Affective Head Set

To test whether our selected head set is sufficiently sparse, we vary K and select the top-K heads according to the ranking determined during refinement in Section 3.2. Figure 3 shows the same trend across all models and datasets: F1 rises quickly with a few heads, saturates near a small K^*, and drops as more heads are added. This suggests that affective signal is concentrated in a small head set, while lower-ranked heads mostly add noise. Since the same K^* appears on both ECC and GoEmotions for each model, the sparsity seems more model-dependent than dataset-specific.

5.2 Selected Heads Attend More to Emotion Words

Setup. We test whether selected heads disproportionately attend to emotion words. For each dialogue turn u that contains at least one word from the NRC Emotion Lexicon (Mohammad and Turney, 2013), let $\mathcal{C}_u$ denote the content-word tokens in the turn, and let $\mathcal{E}_u \subseteq \mathcal{C}_u$ denote the subset of tokens matched by the NRC. For a head h, let $\mathcal{S}_{h,u}$ be the top-30% tokens in $\mathcal{C}_u$ ranked by the attention that head h assigns to them at the prediction position. We then compute a concentration ratio $r_{h,u} = (|\mathcal{S}_{h,u} \cap \mathcal{E}_u|/|\mathcal{S}_{h,u}|)/(|\mathcal{E}_u|/|\mathcal{C}_u|)$. The numerator is the emotion-word fraction among the head's top-attended tokens, while the denominator is the emotion-word fraction in the whole turn. Thus, $r_{h,u} > 1$ indicates that head h attends to emotion words more than expected from their base rate in that turn. We compare selected heads with random non-selected heads sampled from early, middle, and late layer bands (40 heads per band).

Table 2: Emotion-word ratio for selected and random heads. Larger values indicate a higher emotion-word share turn average.

Head set	Gemma3 (K=4)	Qwen3 (K=10)	LLaMA3 (K=20)
Selected	**1.575**	**1.370**	**1.384**
Random Late	1.373	1.064	1.195
Random Middle	1.346	1.157	1.228
Random Early	1.224	1.083	1.092

Table 3: Mean Δlabel_gap per group.

Group	Gemma3	Qwen3	LLaMA3
POS	+9.77	+2.85	+0.79
NEG	−7.57	−2.12	−0.83
HARD	+2.53	+1.14	+0.45

Internal signal is head-specific, not layer-driven. The selected heads have the highest emotion-word ratios for all three models (Table 2). Among random heads, middle and late layer bands generally show higher ratios than early layers, consistent with prior work showing stronger affective or sentiment signals in mid-to-late layers (Tak et al., 2025, Tigges et al., 2024). Moreover, the selected heads span early, middle, and late layer regions, yet outperform random heads within each region. This suggests that the higher emotion-word ratio is a property of the selected heads themselves, not merely a consequence of their layer positions. Appendix F provides qualitative examples.

5.3 What Makes a Sample Rescuable?

SIAB changes the prediction only when its induced logit movement is both large enough and pointed in the useful direction. We first describe the observed movement across POS, NEG, and HARD examples, and then analyze two factors behind it: the baseline logit gap and the alignment of selected heads with the gold label.

SIAB induces different logit movements across groups. We measure the intervention effect by $\Delta\text{label_gap} = \text{label_gap}_{\text{SIAB}} - \text{label_gap}_{\text{ZS}}$. For an example with gold label y, candidate label set

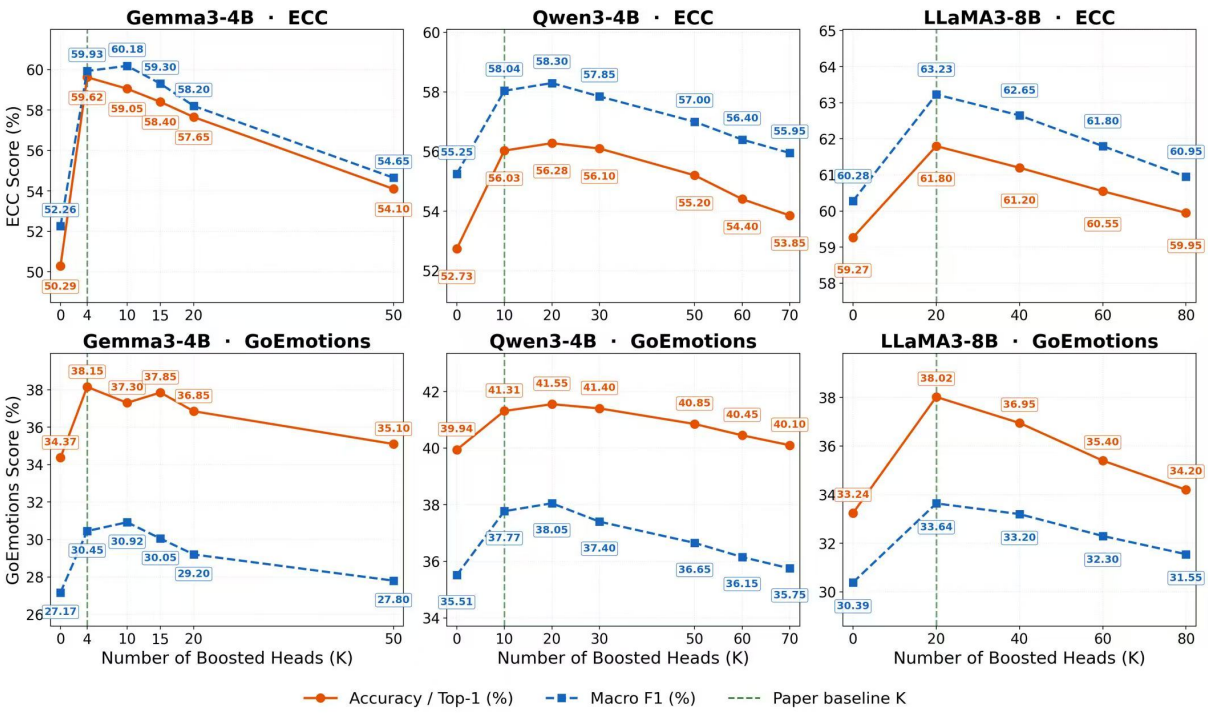

Figure 3: Performance vs. boosted head count K on ECC and GoEmotions. Blue denotes macro-F1; orange denotes accuracy for ECC and top-1 accuracy for GoEmotions. Dashed lines mark the K^* used in the main experiments, showing small-K^* saturation across datasets for each model.

Table 4: (a) ZS margin = $\ell_{\text{top1}} - \ell_{\text{top2}}$ and ZS label_gap = $\ell_{\text{gold}} - \ell_{\text{top1}}$ by group. (b) Rescue rate of POS+HARD samples stratified by ZS label_gap.

(a) Baseline logit geometry by group.

Group	%	ZS margin	ZS label_gap
POS	12.9	9.4	−9.8
NEG	3.1	8.7	0.0
EASY	42.7	20.5	0.0
HARD	41.3	16.8	−20.2

(b) Rescue rate of POS+HARD by ZS label_gap.

Range	%	Rescue rate
$[-5, 0]$ (shallow)	10.2	**66.9%**
$[-10, -5]$ (mid)	12.3	47.9%
$[-100, -10]$ (deep)	77.5	14.4%

Table 5: Selected-head attention mass on emotion tokens. POS shows stronger label-directional bias.

Group	r_{label}	r_{wrong}	Δr
POS	0.767 ± 0.284	0.140 ± 0.233	+0.627
NEG	0.661 ± 0.335	− − −	− − −
EASY	0.726 ± 0.325	− − −	− − −
HARD	0.656 ± 0.345	0.195 ± 0.285	+0.461

$\mathcal{Y}$, and label-logit vector ℓ, we define label_gap $= \ell_y - \max_{c \in \mathcal{Y}} \ell_c$, where c indexes candidate emotion labels. The gap is 0 when the gold label is already top-1 and negative otherwise; positive Δlabel_gap values mean that SIAB moves the gold label closer to top-1.As reported in Table 3, POS shifts strongly upward and NEG strongly downward; HARD also shifts upward, but by a much smaller margin.

Large baseline logit gaps limit rescue. The weaker HARD movement is partly explained by its baseline logit geometry. Table 4 (a) shows that POS and NEG have small top-1 vs. top-2 margins on Gemma3 (9.4 and 8.7), while HARD has a much larger margin (16.8) and a deeper gold-label gap (−20.2). Stratifying POS+HARD examples by ZS label_gap further confirms this pattern (Table 4 (b)): rescue is frequent in the shallow range but drops sharply once the gold label is more than 10 logit points behind top-1. HARD mostly lies in this deep-gap regime, where SIAB can move the gold label but rarely enough to change the prediction.

Direction is set by head alignment. Selected heads differ across groups in how strongly they attend to gold-label versus wrong-label emotion tokens before intervention. Using the NRC EmoLex categories[1], we compute the fraction of selected-

[1]To NRC's eight categories. We map ECC's 10 labels deterministically: *happy / excited / relaxed / proud→joy*,

head attention mass on emotion tokens matching the gold label (r_{label}) versus the wrong top-1 (r_{wrong}). Across all groups, $r_{\text{label}} > 0.65$ (Table 5). POS exhibited stronger directional bias than HARD (0.767 vs. 0.656) and a lower wrong-aligned fraction (0.140 vs. 0.195). On HARD, the intervention thus amplifies a noisier signal, which can partially explain HARD's resistance to rescue.

Can more HARD samples be rescued? Increasing β and suppressing sink-token attention rescue some HARD cases, recovering up to 9% of them when combined. However, both interventions can also disrupt the attention patterns of confident EASY cases. We discuss this in Appendix E.

5.4 Layer-Band Proximity in SIAB–PASTA Additivity

The Results section identified LLaMA3 on ECC as the single non-additive outlier. Here we test whether this outlier is caused by *layer disjointness* between the two head sets.

Quantifying additivity. For each PASTA head set, we compare the actual Combo gain to the sum of the two individual gains as:

$$\text{Additivity} = \frac{\Delta_{\text{Combo}}}{\Delta_{\text{PASTA}} + \Delta_{\text{SIAB}}},$$

where $\Delta_X = X - \text{ZS}$. A value of 100% means the two interventions are exactly additive; $< 100\%$ indicates redundancy or interference between the two head sets and $> 100\%$ indicates super-additivity.

Layer-disjointness hypothesis. We hypothesize that PASTA and SIAB compose additively only when the two head sets occupy overlapping or adjacent Transformer layers, and breaks down when the two sets sit too many layers apart for their signals to combine. LLaMA3 on ECC is the natural witness. Its scan-selected PASTA top-10 lay entirely on $L{\leq}11$, and Combo provided essentially nothing on top of PASTA alone. To probe the causal direction, we (i) re-select LLaMA3's PASTA heads to mid-late layers and check whether additivity recovers, and (ii) artificially reproduce the disjoint pattern in two models where Combo otherwise works by hand-selecting their PASTA-positive heads at near-SIAB vs. early layer bands.

grateful→joy / trust, *sad / disappointed→sadness*, *angry / frustrated→anger*, *worried→fear*.

Table 6: Layer-band manipulation on ECC. *Near* / *Far* denote PASTA heads at small / large mean layer distance $\bar{d}$ to SIAB's heads, respectively. Bold rows mark the high-additivity configuration per model. †Configuration used in the main results (Table 1).

Model	PASTA dist. to SIAB	Additivity (acc / F1)
LLaMA3	Far (native)†	72.7% / 54.3%
LLaMA3	Near (re-selected)	$\mathbf{89.0\%}$ / $\mathbf{108.0\%}$
Gemma3	Near ($\bar{d}$=1.1)	$\mathbf{95.8\%}$ / $\mathbf{96.6\%}$
Gemma3	Far ($\bar{d}$=10.5)	79.4% / 81.5%
Qwen3	Near ($\bar{d}$=0.5)	$\mathbf{118.5\%}$ / $\mathbf{112.7\%}$
Qwen3	Far ($\bar{d}$=11.1)	70.0% / 70.4%

Manipulating PASTA's layer band recovers and breaks additivity. Table 6 reports additivity per manipulation, and the pattern is consistent across all three models. Additivity is high when PASTA and SIAB occupy nearby layers and collapses when they are forced apart. Re-selecting LLaMA3's natively early-stacked PASTA heads ($\bar{d}$=8.6) to mid-late layers ($\bar{d}$=0.3) lifted F1 additivity above 100% (108.0%). Conversely, forcing the otherwise-additive Gemma3 and Qwen3 PASTA heads into early layers dropped additivity from 95.8% and 118.5% (near) to 79.4% and 70.0% (far). The recovery and breakdown both track the layer gap rather than the model, confirming that the layer band is the limiting factor.

6 Conclusion

Affective state in dialogue depends on contextual cues that are already present but easily lost at prediction time. We show that SIAB can route this evidence through internal model signals, without relying on externally marked spans or task-specific training. A fixed causally localized head set transfers across two contrasting attentional demands, whereas PASTA-style external profiles are competitive in-domain but transfer poorly. Their additive gains further suggest that internal and external signals capture complementary aspects of affective evidence. Future work should move from a fixed global intervention to sample-adaptive control, where intervention strength and target tokens vary with each example's uncertainty and attention pattern, and test whether the same interface extends to other latent conversational states such as uncertainty, trust, or intent.

Limitations

Model scale. We evaluated three open instruction-tuned LLMs spanning 4–8B parameters from distinct families (Gemma3, Qwen3, LLaMA3). Whether the internal signal continues to generalize at much larger scales (e.g., 70B+) is an open question Exhaustive activation patching at such scales is computationally prohibitive in our setting.

Language coverage. Both datasets (ECC and GoEmotions) were in English. Whether the recovered head set transfers across languages, or whether language-specific recognition circuits would emerge under multilingual evaluation, is untested.

Dataset breadth. The two datasets were selected to span complementary attentional demands (long-context dialogue integration vs. wide-label distribution shift over short inputs), and we reuse the same head set across them without re-derivation. Generalization to non-emotion classification tasks is left to future work.

Boundary on HARD samples. A subset of test samples (HARD) remained unrescued even under stronger interventions. As analyzed in Appendix E, scaling β alone is insufficient, and global sink suppression flips confident EASY samples in similar magnitude, yielding no net gain. Whether these samples can be rescued by a more selective, sample-adaptive intervention is left to future work.

References

Nora Belrose, Zach Furman, Logan Smith, Danny Halawi, Igor Ostrovsky, et al. March 2023. Eliciting Latent Predictions from Transformers with the Tuned Lens. *Computing Research Repository, arXiv Preprint,* arXiv:2303.08112.

Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, et al. July 2020. GoEmotions: A Dataset of Fine-grained Emotions. *In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4040–4054, Online.

Paul Ekman. 1992. An Argument for Basic Emotions. *Cognition & Emotion*, 6(3–4):169–200.

Gino Franco Fazzi, Julie Skoven Hinge, Stefan Heinrich, and Paolo Burelli. April 2025. Don't Get Too Excited —Eliciting Emotions in LLMs. *In Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, pages 200:1–200:9, New York, NY, USA.

Asbjørn Følstad, Effie L.-C. Law, and Nena van As. November 2024. Conversational Breakdown in a Customer Service Chatbot: Impact of Task Order and Criticality on User Trust and Emotion. *ACM Transactions on Computer-Human Interaction*, 31(5):1–52.

Yumeng Fu, Junjie Wu, Zhongjie Wang, Meishan Zhang, Lili Shan, et al. January 2025. LaERC-S: Improving LLM-based Emotion Recognition in Conversation with Speaker Characteristics. *In Proceedings of the 31st International Conference on Computational Linguistics*, pages 6748–6761, Abu Dhabi, UAE.

Gemma Team. March 2025. Gemma 3 Technical Report. *Computing Research Repository, arXiv Preprint,* arXiv:2503.19786.

Nicholas Goldowsky-Dill, Chris MacLeod, Lucas Sato, and Aryaman Arora. April 2023. Localizing Model Behavior with Path Patching. *Computing Research Repository, arXiv Preprint,* arXiv:2304.05969.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, et al. July 2024. The Llama 3 Herd of Models. *Computing Research Repository, arXiv Preprint,* arXiv:2407.21783.

Yuanyuan He, Yongsen Pan, Wei Li, Jiali You, Jiawen Deng, et al. November 2025. ECC: An Emotion-Cause Conversation Dataset for Empathy Response. *In Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 6000–6017, Suzhou, Jiangsu, China.

Goro Kobayashi, Tatsuki Kuribayashi, Sho Yokoi, and Kentaro Inui. November 2020. Attention is Not Only a Weight: Analyzing Transformers with Vector Norms. *In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 7057–7075, Online.

Shanglin Lei, Guanting Dong, Xiaoping Wang, Keheng Wang, Runqi Qiao, et al. September 2023. InstructERC: Reforming Emotion Recognition in Conversation with Multi-task Retrieval-augmented Large Language Models. *Computing Research Repository, arXiv Preprint,* arXiv:2309.11911.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. 2023. Inference-Time Intervention: Eliciting Truthful Answers from a Language Model. *Advances in Neural Information Processing Systems,* 36:41451–41530.

Xinran Li, Yu Liu, Jiaqi Qiao, and Xiujuan Xu. March 2026. Do LLMs Feel? Teaching Emotion Recognition with Prompts, Retrieval, and Curriculum Learning. *In Proceedings of the 40th AAAI Conference on Artificial Intelligence,* pages 31778–31786, Singapore.

Paul Michel, Omer Levy, and Graham Neubig. 2019. Are Sixteen Heads Really Better than One? *In Proceedings of the 33rd International Conference on Neural Information Processing Systems,* 1257:14037–14047, Red Hook, NY, USA.

Saif M. Mohammad and Peter D. Turney. 2013. Crowdsourcing a Word–Emotion Association Lexicon. *Computational Intelligence*, 29(3):436–465.

Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, et al. 2022. In-context Learning and Induction Heads. *Computing Research Repository, arXiv Preprint,* arXiv:2209.11895.

OpenAI. March 2026. GPT-5.4 thinking system card. Technical report, OpenAI.

Nina Rimsky, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, et al. August 2024. Steering Llama 2 via Contrastive Activation Addition. *In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15504–15522, Bangkok, Thailand.

Ala N. Tak, Amin Banayeeanzade, Anahita Bolourani, Mina Kian, Robin Jia, et al. July 2025. Mechanistic Interpretability of Emotion Inference in Large Language Models. *Findings of the Association for Computational Linguistics: ACL 2025*, pages 13090–13120, Vienna, Austria.

Curt Tigges, Oskar J. Hollinsworth, Atticus Geiger, and Neel Nanda. November 2024. Language Models Linearly Represent Sentiment. *In Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pages 58–87, Miami, FL, USA.

Curt Tigges, Oskar John Hollinsworth, Atticus Geiger, and Neel Nanda. October 2023. Linear Representations of Sentiment in Large Language Models. *Computing Research Repository, arXiv Preprint,* arXiv:2310.15154.

Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J. Vazquez, et al. 2025. Steering Language Models with Activation Engineering. *In Proceedings of the 13th International Conference on Learning Representations,* 14 pages.

Elena Voita, David Talbot, Fedor Moiseev, Rico Sennrich, and Ivan Titov. July 2019. Analyzing Multi-Head Self-attention: Specialized Heads Do the Heavy Lifting, the Rest Can Be Pruned. *In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5797–5808, Florence, Tuscany, Italy.

Chenxi Wang, Yixuan Zhang, Ruiji Yu, Yufei Zheng, Lang Gao, et al. October 2025. Do LLMs "Feel"? Emotion Circuits Discovery and Control. *Computing Research Repository, arXiv Preprint,* arXiv:2510.11328.

Kevin Ro Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. 2023. Interpretability in the Wild: A Circuit for Indirect Object Identification in GPT-2 Small. *In Proceedings of the 11th International Conference on Learning,* 21 pages.

Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. 2024. Efficient Streaming Language Models with Attention Sinks. *In Proceedings of the 12th International Conference on Learning Representations,* 21 pages.

Jieying Xue, Minh-Phuong Nguyen, Blake Matheny, and Le-Minh Nguyen. 2024. BiosERC: Integrating Biography Speakers Supported by LLMs for ERC Tasks. *In Proceedings of the 33rd International Conference on Artificial Neural Networks,* Part V, pages 277–292, Lugano, Switzerland.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, et al. May 2025. Qwen3 Technical Report. *Computing Research Repository, arXiv Preprint,* arXiv:2505.09388.

Zhongzhi Yu, Zheng Wang, Yonggan Fu, Huihong Shi, Khalid Shaikh, et al. 2024. Unveiling and Harnessing Hidden Attention Sinks: Enhancing Large Language Models without Training through Attention Calibration. *In Proceedings of the 41st International Conference on Machine Learning*, pages 57659–57677, Vienna, Austria.

Qingru Zhang, Chandan Singh, Liyuan Liu, Xiaodong Liu, Bin Yu, et al. 2024a. Tell Your Model Where to Attend: Post-hoc Attention Steering for LLMs. *In Proceedings of the 12th International Conference on Learning Representations,* 20 pages.

Qingru Zhang, Xiaodong Yu, Chandan Singh, Xiaodong Liu, Liyuan Liu, et al. 2024b. Model Tells Itself Where to Attend: Faithfulness Meets Automatic Attention Steering. *Computing Research Repository, arXiv Preprint,* arXiv:2409.10790.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, et al. October 2023. Representation Engineering: A Top-down Approach to AI Transparency. *Computing Research Repository, arXiv Preprint,* arXiv:2310.01405.

A Pair Construction for Causal Patching

Following the minimal-pair design of Wang et al. (2023), each pair $(X^{\text{clean}}, X^{\text{corr}})$ shares syntactic skeleton and token length and differs only in affective content.

Coverage. The set spans the six basic emotions of Ekman (1992): anger, disgust, fear, happiness, sadness and surprise, 100 pairs each (600 total), covering both valences and arousal levels to avoid a polarity confound. Example (anger): clean "`During class, you blamed me for the result, and I feel very`" (target `angry`) vs. corrupted "`. . . you praised me. . .`" (target `happy`). Within each category, pairs further vary the syntactic locus of the cue —verb, predicate-adjective, prenominal-adjective, or purpose-clause

replacement— and distribute it across multiple content words, so the discovered head set is not tied to a single syntactic position.

Verification. Pairs were generated by GPT-5.4 (OpenAI, 2026) (distinct from the downstream models) and human-verified for grammaticality, token-length match, syntactic skeleton preservation, and target high-rank continuation.

Generalization beyond localization categories. Only 4 of 10 ECC labels and 6 of 27 GoEmotions labels overlap with the pair-set categories; the rest are unseen during localization. Strong performance on these unseen labels (§4.5) indicates that the head set generalizes beyond the specific cues used to identify it.

B Implementation Details for Refinement and Intervention

Table 7 summarizes the complete set of hyperparameters used by SIAB, listing the role, value, and how each is set; the per-model final head set $\mathcal{H}$ and the calibrated intervention strength β are reported in Table 8.

Non-content token filter and heads detail. A token is excluded from A_h, V_h averaging and from the intervention pool $\mathcal{T}_r(h, q)$ if it falls into any of: (i) punctuation, (ii) English personal pronouns, (iii) interrogatives (*what, who, when, where, why, how, which*), (iv) prepositions, or (v) model-specific special tokens (BOS / EOS and chat-template markers). Negation and intensifiers (*not, very, no, . . .*) are retained as content tokens.

C Hyperparameters and Calibration

This section describes how the calibration sets are constructed and how each hyperparameter is tuned or frozen; the GoEmotions multi-label decoding parameters (T, t) used for each (model, shot) configuration are listed in Table 9.

Calibration sets. For ECC, 200 dialogues sampled from the official dev split form the calibration set; evaluation uses the full ECC test split. For GoEmotions, 200 dev samples form the calibration set; evaluation uses the full GoEmotions test split, excluding Neutral-only samples. The per-model β is calibrated once on the ECC dev set and reused unchanged on GoEmotions. No hyperparameter is tuned on test data.

Hyperparameter inventory. The structural hyperparameters ($\tau_{\text{head}}, \lambda, K, |\mathcal{Q}|andr$) are calibrated once on the ECC dev set and frozen across all (model, dataset, shot) configurations. The bias magnitude β is likewise calibrated once per model on the ECC dev set and held fixed across both datasets. $|\mathcal{Q}|$=4 tokens around the answer position; r=0.3 of content-word tokens; K=4, 10, 20 for Gemma3, Qwen3, LLaMA3 respectively, set at each model's saturation point on the ECC dev sweep (Figure 3); β=6, 9, 10 for Gemma3, Qwen3, LLaMA3 respectively. The GoEmotions multi-label decoding uses a top-T cap and inclusion threshold t, tuned on the GoEmotions dev set under the baseline and frozen for +SIAB; per-(model, shot) values are in Table 9.

D Results on Neutral-Only GoEmotions Samples

We exclude samples labeled solely as Neutral, since they lack explicit emotion cues. Multi-label samples retain their non-Neutral labels. On the excluded Neutral-only subset (1,606 test samples), SIAB *improved* Neutral recall by $+1.08$ to $+1.93$ percentage points across the three models (Gemma3: 9.10 $\rightarrow$ 10.54; Qwen3: 10.05 $\rightarrow$ 11.13; LLaMA3: 8.93 $\rightarrow$ 10.86), indicating that content-token amplification does not inflate emotion predictions on neutral inputs.

E Why HARD Samples Resist Intervention

Can stronger or alternative interventions shift HARD samples into the rescuable regime? We probe two perturbations —scaling β and suppressing sink-token attention on selected heads (Table 10).

Scaling β alone is insufficient. Doubling β rescued only 4.5% of HARD samples on the Gemma3 model and saturated beyond $\times 2$, the LLaMA3 and Qwen3 models yielded essentially no gain (0.1% and 0.4%, Δlabel_gap ≈ 0). The HARD ceiling is not magnitude-bounded.

Attention sinks absorb selected-head capacity. Structural tokens (BOS, EOS, punctuation) absorb $\sim$25%, $\sim$40%, and $\sim$60% of top-1 mass on Gemma3, Qwen3, and LLaMA3 —a known attention sink phenomenon (Xiao et al., 2024)— and similar levels on EASY samples, so sink dominance is a head-level property rather than a HARD-specific failure. Zeroing sink attention recovered an

Table 7: Hyperparameter inventory. The first six are fixed per model across all (dataset, shot) configurations, while (T, t) is recalibrated as noted.

Symbol	Role	Value	How set
τ_{head}	Direct-effect threshold	0.05	(Wang et al., 2023)
λ	Attention-vs.-value weight in s_{head}	0.5	Equal-weighting
K	Heads in $\mathcal{H}$	4, 10, 20	K^* saturation, Fig. 3
$\lvert\mathcal{Q}\rvert$	Boost-applied query positions	4	ECC dev
r	Top fraction of attended content tokens	0.3	A priori
β	Pre-softmax bias magnitude	Tab 8	Per model
(T, t)	GoEmotions multi-label top-T + threshold	Tab 9	Per (model, shot), baseline only

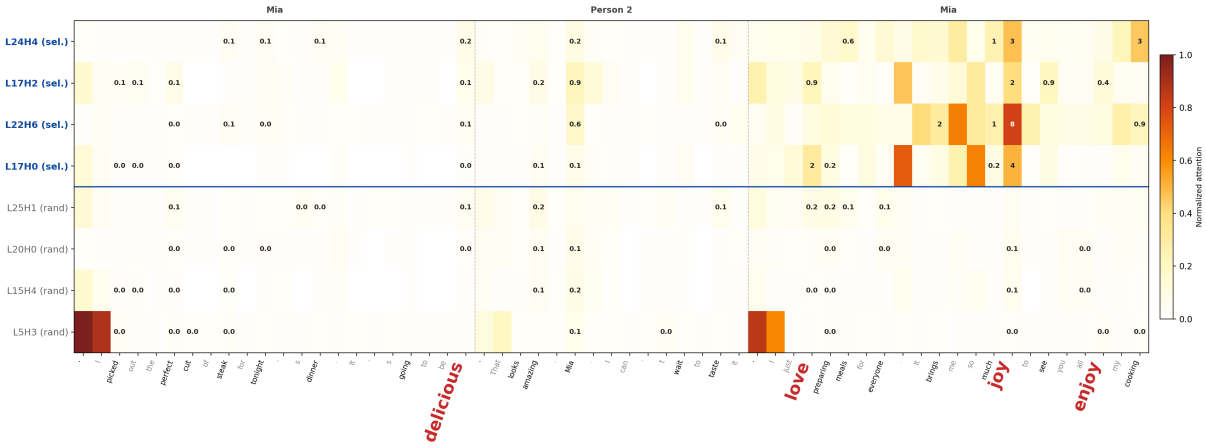

Figure 4: Per-head attention on an ECC example rescued by SIAB (GT *happy*; ZS $\rightarrow$ *proud* ✗, ZS+SIAB $\rightarrow$ *happy* ✓). Each row: attention from the final-4 query band on a single head, globally normalized; cell value = attention $\times 100$, restricted to the head's top-30% content tokens (the boost set $\mathcal{S}_{h,q}$). Bold red tokens are NRC EmoLex emotion words. Top: the four Gemma3 selected heads (L24H4, L17H2, L22H6, L17H0). Bottom: depth-stratified random non-selected heads (L25H1, L20H0, L15H4, L5H3).

additional 4.2% HARD samples on Gemma3 (combined 8.7%, Δlabel_gap from $+0.82$ to $+5.89$) and 5.5% on LLaMA3, showing that sinks lock capacity magnitude scaling cannot reach. Qwen3 was the exception (1.4%); several heads concentrated $> 60\%$ of attention on a single sink token, and zeroing destabilized the distribution— sinks here were integral to head behavior, not a removable artifact.

Sink suppression is not deployable. Applied globally, β+sink suppression raised HARD rescue but flipped EASY samples in similar magnitude, yielding no net F1 gain. Sinks acted as a barrier on confused baselines but as a stabilizer on confident ones, and EASY / HARD was unobservable at inference. We thus keep β-only as default and report sink suppression as a diagnostic of the structural limit.

F Qualitative Attention Visualization

Figure 4 visualizes per-head attention on an ECC dataset example, contrasting the four selected heads (top) from the Gemma3 model with four depth-stratified random non-selected heads (bottom).

G Prompt

Zero-shot. Inputs of the ECC dataset are rendered as a multi-turn chat (user / assistant alternation along dialogue turns, assistant role assigned to the target speaker) via `tokenizer.apply_chat_template`. A final user message asks for the emotion of the last utterance with the 10-class label list. The inputs of the GoEmotions dataset use a single user message containing the sentence and the 27-class label list. System message is the default "You are a helpful assistant" for both. The ECC final-turn instruction is:

Table 8: Final head set $\mathcal{H}$ per model, with intervention strength β. Both $\mathcal{H}$ and β are fixed per model and reused across the two datasets. All indices are 0-based. HuggingFace identifiers: `google/gemma-3-4b-it`, `Qwen/Qwen3-4B-Instruct-2507`, `meta-llama/Llama-3.1-8B-Instruct`.

Model	Selected (layer, head) indices
Gemma3 (K=4, β=6)	(24,4), (17,2), (22,6), (17,0)
Qwen3 (K=10, β=9)	(22,5), (18,30), (23,10), (20,9), (29,3), (26,27), (29,0), (7,3), (25,9), (24,15)
LLaMA3 (K=20, β=10)	(18,0), (16,10), (19,9), (16,26), (21,26), (6,20), (22,1), (13,10), (15,26), (21,24), (15,14), (19,11), (14,24), (15,20), (20,9), (13,12), (17,28), (24,21), (28,25), (17,20)

Table 9: GoEmotions multi-label decoding parameters (T, t) per (model, shot) pair: T = Temperature, t = inclusion threshold. Tuned on GoEmotions dev under baseline and frozen for +SIAB, so reported ΔF1 reflects intervention effect, not (T, t) re-calibration.

Shots	Qwen3	LLaMA3	Gemma3
0	(5, 0.10)	(3, 0.10)	(6, 0.05)
5	(3, 0.20)	(3, 0.10)	(7, 0.05)
10	(2, 0.20)	(3, 0.10)	(10, 0.05)
15	(7, 0.08)	(3, 0.10)	(3, 0.03)
27	(3, 0.15)	(2, 0.15)	(7, 0.05)

```
What emotion is expressed in the last
    utterance?
Answer with one word: happy, excited, sad,
    proud, frustrated,
angry, relaxed, afraid, confident, or
    guilty.
```

Few-shot. Demonstrations follow an *utterance / cues / reasoning / emotion* chain-of-thought format, with one demonstration per emotion class (10 for ECC and 27 for GoEmotions). Demonstrations are prepended to the zero-shot prompt and held fixed across methods within each (dataset, shot) setting. One ECC demonstration is shown below as a template:

```
Utterance: "I can't believe I finally got
    the promotion after
all these years!"
Cues: "can't believe", "finally", "got the
    promotion"
   -> excitement from achieving a
    long-awaited goal
Emotion: excited
```

Table 10: Rescue rate on HARD samples under three intervention strategies. *Sink mass*: Mean Top-1 attention mass on selected heads, dominated by sink tokens (BOS/EOS/punctuation). *Increase β*: Additional rescue from doubling β alone over the model's baseline. *Combined*: Rescue from doubling β together with full sink suppression.

Model	Sink mass	Increase β	Sink suppress.	Combined β + sink
Gemma3	∼25%	4.5%	− − −	**8.7%**
LLaMA3	∼60%	0.1%	− − −	**5.5%**
Qwen3	∼40%	0.4%	− − −	1.4%

Prompt for Refinement Dialogue Generation The 200 dialogues used for label-free head refinement (§3.2) are generated by GPT-5.4 (OpenAI, 2026) with the prompt below. The prompt requests two-speaker dialogues whose final-turn emotion is inferable from cross-turn context rather than a single explicit emotion word, balanced across the six basic emotions of Ekman (1992), matching the coverage of the patching pairs (Appendix A). The per-sample `target_emotion` and `rationale` fields are stored only as metadata; refinement uses neither, consistent with its label-free design (§3.2).

```
Prompt for generating synthetic
    emotion-push dialogues.
Generate complete two-speaker dialogues
    between Speaker A and Speaker B.
Each sample should be an independent
    dialogue. The goal is to create
    dialogues in which the emotional state
    of the
target speaker in the final turn can be
    inferred from the preceding context.
Requirements:
1. Each dialogue must contain 6 to 12 turns
    in total.
2. The final turn must be spoken by the
    target speaker. The target speaker's
   emotional state in the final turn must
    belong to one of the six basic
   emotions proposed by Ekman (1992):
   anger, disgust, fear, happiness,
    sadness, or surprise.
3. The target emotion should not be
    expressed only through an explicit
    emotion
   word such as "I am angry", "I am sad",
    or "I am afraid". Instead, the emotion
   must be inferable from the dialogue
    context, including the speaker's
    situation,
   expectations, reactions, and changes
    across turns.
4. The emotional cue must be cross-turn and
    reasoning-based. Do not make the
   emotion depend on a single isolated
    keyword. The dialogue should contain
```

```
   earlier contextual evidence that becomes
     relevant to the target speaker's
   final emotional state.
5. The dialogue should include a natural
     conversational transition. The topic may
   shift during the dialogue, but the
     earlier affective cue should remain
     relevant
   to interpreting the final turn.
6. Avoid artificial or overly dramatic
     writing. The dialogue should sound like
     a
   plausible everyday conversation between
     two people.
7. Do not include the emotion label inside
     the dialogue. Provide the label only
   in the metadata.
8. Balance the generated samples across the
     six emotions. For each emotion,
   generate the same number of dialogues.
9. Use the following output format for each
     sample:
{
  "id": "...",
  "target_emotion":
     "anger/disgust/fear/happiness
  /sadness/surprise",
  "target_speaker": "A or B",
  "dialogue": [
    {"speaker": "A", "utterance": "..."},
    {"speaker": "B", "utterance": "..."}
  ],
  "rationale": "Briefly explain which
     cross-turn cues support the target
     emotion."
}
```