Some Emotions Run Deeper: Layer-wise Probing and Causal Intervention in Large Language Models

Anonymous ACL submission

Abstract

Emotion is expressed in text along a wide spectrum, from surface lexical cues to inferences entangled with content. Most layer-wise analyses of emotion in LLMs use a single corpus, leaving open whether the depth at which emotion becomes accessible is a property of the model or also of the text source. We investigate this across three datasets spanning different degrees of explicitness and contextualization in emotion expression (Twitter posts, Reddit comments, and autobiographical narratives) and eight 1B–9B open-weight LLMs from the Llama, Qwen, and Granite families. We combine layer-wise probing with offline and online causal interventions, transfer analyses, and an early-exit classifier. We find that (i) the best probing layer shifts systematically across text sources, from input-adjacent layers to over half model depth, and the surrounding high-performance band broadens from $\sim$2 to over 10 layers; (ii) forward-pass interventions on probe-selected bands reduce test accuracy by 5–6 points more than same-width random bands ($q < 0.01$); (iii) selected bands transfer across datasets and emotion categories, suggesting partially shared affective information rather than strictly per-emotion substrates; and (iv) probe-selected early-exit representations outperform full-depth exits by 6.9 points and exceed frozen RoBERTa and DeBERTa baselines.

1 Introduction

Large language models are increasingly used in affect-sensitive settings, including emotion classification, emotional support, and broader tests of emotional intelligence (Sabour et al., 2024; Wang et al., 2023; Kang et al., 2024). Although prior evaluations show that LLMs can recognize and respond to emotional content at the output level, less is known about how affective information is organized internally: whether it becomes accessible at a fixed depth in decoder-only LLMs, or varies with the kind of text being processed.

This question matters because emotion is expressed differently across text sources. In short social-media posts, affect often appears through surface cues such as lexical markers, emotion hashtags, emoji/emoticons, punctuation, or conventional phrases (Mohammad and Kiritchenko, 2015; Mohammad and Bravo-Marquez, 2017; Dresner and Herring, 2010). In longer comments and autobiographical narratives, it may instead be conveyed implicitly through situations, appraisals, causal structure, or pragmatic inference (Balahur et al., 2011; Hofmann et al., 2020; Troiano et al., 2023; Scherer and Wallbott, 1994a).

Layer-wise probing is a natural tool for locating where emotion becomes accessible in LLM representations (Alain and Bengio, 2017; Belinkov, 2022). Prior work shows that linguistic, semantic, and factual information peak at different transformer depths (Tenney et al., 2019; Rogers et al., 2020; Geva et al., 2023), and that sentiment and emotion are most detectable in early-to-mid Llama layers (Di Palma et al., 2025). Whether this affective pattern generalizes across model families and text sources remains unclear. Because decodability alone does not imply functional relevance (Hewitt and Liang, 2019; Belinkov, 2022), we pair probing with intervention-based perturbations and early-exit readout (Fan et al., 2020; Sajjad et al., 2023; McGrath et al., 2023; Xin et al., 2020).

We study where emotion-sensitive information becomes accessible in decoder-only LLMs, and whether probe-selected layers are merely diagnostic or also useful for intervention and readout. We analyze eight open-weight LLMs from three families (Llama, Qwen, and Granite; 1B–9B parameters) on three stylistically different English emotion corpora: *Emotion*/CARER, *GoEmotions*, and *ISEAR*. We restrict them to the shared labels fear, joy, anger, and sadness, use frozen representations

with fixed deterministic pooling and linear probes, and evaluate selected layer bands through online intervention, offline feature scaling, transfer analyses, and early-exit classification. Figure 1 summarizes the pipeline.

Our main findings are fourfold:

- Emotion depth varies systematically with dataset source: Emotion/CARER peaks near input-adjacent layers, GoEmotions is intermediate and diffuse, and ISEAR peaks much deeper.
- Probe-selected bands are causally engaged in the probe-mediated readout, as online interventions disrupt prediction more than same-width unrelated controls.
- Transfer depends on the evaluation setting, with stronger sharing under intervention than under early-exit readout.
- Probe-selected bands support compact readout: selected exits outperform full-depth, random, low-sensitivity, transfer, and frozen-encoder baselines under the same linear protocol.

Together, these results show that affective information is localized enough to be causally disrupted and efficiently exploited, but its depth and generalizability are not universal: they depend on what kind of text is being processed, which model family is used, and how the representation is decoded.

2 Related Work

Our work draws on three lines of prior research: emotion understanding in large language models, layer-wise probing of model representations, and layer-level interventions together with their use for efficient inference.

Emotion Understanding in Language Models. Discrete emotion categorization is a long-standing NLP task, grounded in psychological theories of basic emotions (Ekman, 1992; Plutchik, 1980). A growing body of work has evaluated large language models on emotion classification across narrative, social-media, and dialogue domains (e.g., Wang et al., 2023; Sabour et al., 2024), generally reporting strong end-to-end performance. This literature characterizes what an LLM predicts rather than how the prediction arises from internal representations. We complement these benchmark-driven results with a representation-level probing analysis of decoder-only LLMs, treating internal hidden states as the unit of study.

Layer-wise Probing. Probing classifiers (Alain and Bengio, 2017; Belinkov, 2022) have become a standard methodology for analyzing information encoded in intermediate representations of neural language models. Hewitt and Liang (2019) argue that probes should be as simple as possible, typically linear, so that findings reflect representation quality rather than probe capacity. Layer-wise analyses on BERT-style models have consistently found that lower layers tend to capture surface and syntactic features while higher layers encode more abstract semantic information (Tenney et al., 2019; Rogers et al., 2020). Related forms of depth-structured processing have also been observed for factual recall in decoder-only language models (Geva et al., 2023). Beyond linguistic and factual content, probing has also been used to localize numerical reasoning (Wallace et al., 2019), spatial and temporal knowledge (Gurnee and Tegmark, 2024), and high-level concept directions such as truthfulness in instruction-tuned models (Marks and Tegmark, 2024). The most directly related work, Di Palma et al. (2025), probes sentiment and emotion in the Llama family to identify an effective probe-and-pooling combination, reporting that affective information is most detectable in the early-to-mid layers. Our focus is on how emotion is organized across multiple model families, datasets, and individual emotion categories, validated through causal interventions.

Layer-level Interventions and Efficient Inference. While probing identifies what information is decodable from representations, it does not establish whether that information plays a causal role in model behavior. A complementary line of work tests causal contribution by intervening on internal representations. Layer-level interventions, which ablate, drop, or perturb entire layers and measure the behavioral consequence (Fan et al., 2020; Sajjad et al., 2023; McGrath et al., 2023), are particularly suited to questions about whether information at a given depth is causally required for a downstream task.

A second line of work leverages the layer-organized structure of transformers for inference-time efficiency. Early-exit methods (Schwartz et al., 2020; Xin et al., 2020; Schuster et al., 2022) learn policies that halt computation at an intermediate layer when an input is sufficiently easy, trading accuracy for compute. In the emotion setting, Di Palma et al. (2025) apply this idea to Llama via SENTRILLAMA, which truncates the model at

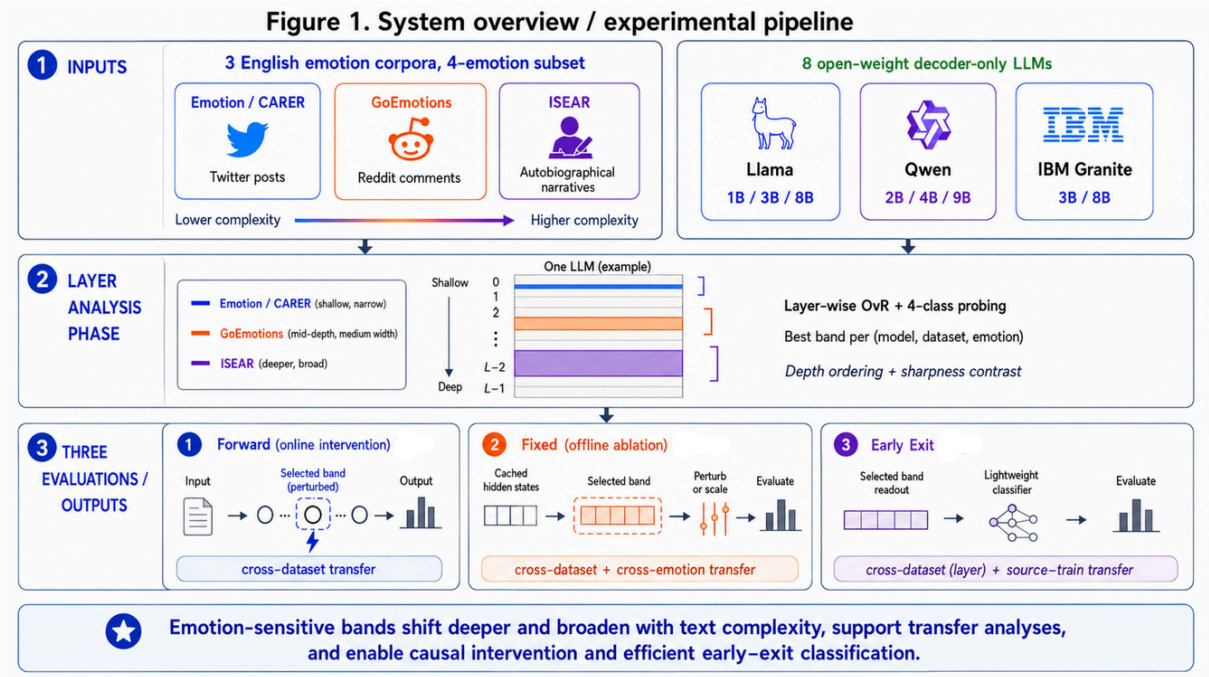

Figure 1: Overview of our experimental pipeline. The term "complexity" is used only as a descriptive shorthand for source/style differences, from short texts with more surface lexical cues to longer, more contextualized narrative texts. We localize emotion-sensitive layer bands with layer-wise probing, then evaluate them through online intervention, offline feature scaling, transfer analyses, and early-exit classification.

a single probe-identified layer. We extend this direction by comparing layer bands of varying width to test whether a contiguous multi-layer band outperforms single-layer selection, and by evaluating cross-dataset transfer of selected bands.

3 Preliminaries

Datasets. We evaluate on three publicly available emotion-classification corpora loaded from HuggingFace: **Emotion** (Saravia et al., 2018), Twitter posts labeled with six emotions; **ISEAR** (Scherer and Wallbott, 1994b), autobiographical narratives elicited via emotion-recall questionnaires across seven emotional categories; and **GoEmotions** (Demszky et al., 2020), Reddit comments labeled with 27 fine-grained emotions plus neutral. From each corpus, we retain the shared target labels fear, joy, anger, and sadness, and discard examples shorter than six whitespace tokens. For GoEmotions, we additionally keep only single-label examples whose sole label is one of the four target emotions. We then construct stratified splits with seed 42: 80/10/10 for Emotion and ISEAR, and 70/15/15 for the resulting smaller GoEmotions subset. Emotion is capped at 1,500 examples per emotion before splitting. Table 1 gives per-split sizes; per-emotion counts are reported in Appendix A.

Dataset	Train	Val	Test	Total
Emotion (CARER)	4,800	600	600	6,000
ISEAR	3,234	402	408	4,044
GoEmotions	2,147	459	463	3,069

Table 1: Dataset sizes after restriction to the four-emotion subset and stratified splitting (80/10/10 for Emotion and ISEAR; 70/15/15 for GoEmotions).

Models. We probe eight open-weight decoder-only large language models from HuggingFace across three families:

- **Llama**: Llama-3.2-1B-Instruct, Llama-3.2-3B-Instruct, Llama-3.1-8B-Instruct (Grattafiori et al., 2024)
- **Qwen**: Qwen-3.5-2B, Qwen-3.5-4B, Qwen-3.5-9B (Yang et al., 2025)
- **Granite**: Granite-4.1-3B, Granite-4.1-8B (IBM Research, 2026)

Pooling and Probing. Our work does not aim to identify the optimal pooling function or probe classifier for emotion probing, a question recently studied at length by Di Palma et al. (2025). We restrict the pooling function to fixed, deterministic operations that apply identically across all models, layers, and datasets. As verified in Appendix B, a learnable single-head attention pooling

baseline exhibits severe training instability at the random seed used for our main experiments, with multiple (layer, emotion) cells failing to converge. To avoid a combinatorial sweep over the remaining fixed options on every model and dataset, we adopt a greedy two-step selection on the validation split: first comparing last-token, mean, and `concat(mean,max,min)` pooling with a linear SVM fixed on Llama-3.2-3B-Instruct, then comparing linear SVM, logistic regression, and MLP probes with `concat` fixed. `concat` pooling with logistic regression is the strongest and most stable choice; the same ordering replicates on Qwen-3.5-4B as a cross-family sanity check. All main experiments therefore use a **StandardScaler + LogisticRegression** probe (C=1.0, `max_iter`= 5000, seed = 42) on top of `concat(mean, max, min)` pooling. For each target emotion we train a per-layer one-vs-rest binary probe (target emotion versus the other three target emotions) on the training split.

4 Where and How do emotions reside in LLMs?

Before intervention, we first localize where emotion information is linearly decodable in frozen LLM representations. For each dataset, model, target emotion, and layer, we train a probe on the training split and evaluate it on validation only. Because models have different depths, we report normalized depth, $(\ell + 1)/L$, where ℓ is the zero-indexed layer and L is the total number of layers. This analysis is descriptive, not causal: high validation F1 indicates decodability, not that the layer causes emotional behavior. We test causal relevance later through fixed and online interventions.

Dataset source style shifts emotion-cue depth. Figure 2 shows normalized-depth profiles across datasets and model families. The main pattern is dataset-driven. Emotion peaks very early, with mean best relative depth 0.066 and median depth 0.036; GoEmotions is more intermediate and diffuse, with mean best depth 0.219; and ISEAR peaks much later, with mean and median depths 0.590 and 0.598. Thus, short social-media-style posts expose emotion cues early, whereas autobiographical ISEAR narratives require deeper processing.

Peak shape also differs across datasets. Figure 3 shows mean validation F1 with IQR shading over normalized depth. Emotion is early and sharply concentrated: layers within 0.01 F1 of the peak span only 0.073 normalized depth on average, or about 2.1 raw layers. GoEmotions is much more diffuse (0.316 depth; 9.0 layers), while ISEAR is later and moderately broad (0.193 depth; 5.9 layers). Thus, datasets differ not only in peak location, but also in signal concentration.

Emotions differ in difficulty and stability. Joy is the most consistently decodable emotion, ranking first in all three datasets and in $19/24$ dataset–model contexts. Sadness is usually hardest, ranking last on Emotion and ISEAR and in $17/24$ dataset–model contexts, although it rises to second on GoEmotions. Fear is competitive on Emotion and ISEAR but weakest on GoEmotions, while anger is usually intermediate.

OvR patterns hold under a four-class probe. Because later interventions require target-emotion-specific bands, our main layer-wise analysis uses one-vs-rest probes. As a sanity check, we also train a four-class multinomial probe at each layer. Across 24 dataset–model pairs, four-class best depth strongly agrees with the mean OvR best depth ($r = 0.85$), and 87.5% of four-class best layers fall within the range of the four emotion-specific OvR best layers. Full results are in Appendix C.

Implications for intervention. These results motivate the intervention design in the next sections. First, because best depth varies systematically by dataset, layer selection must be performed on validation data within each dataset rather than using a universal raw layer index. Second, because the high-F1 region can be narrow or broad depending on the dataset, we test multiple band widths rather than a single best layer only. Third, because emotion rankings are stable but not identical across datasets, transfer experiments are necessary: if emotion-sensitive regions capture general emotional information, then bands selected for one emotion or dataset should have measurable effects beyond their original selection context.

5 Causal Intervention on Emotion-Sensitive Layer Bands

The layer-wise analysis (§4) shows that emotion is linearly decodable at dataset-dependent depths. We next ask whether these regions matter for downstream readout. We test this with two complementary interventions: an *online* intervention that scales activations during the forward pass, and an

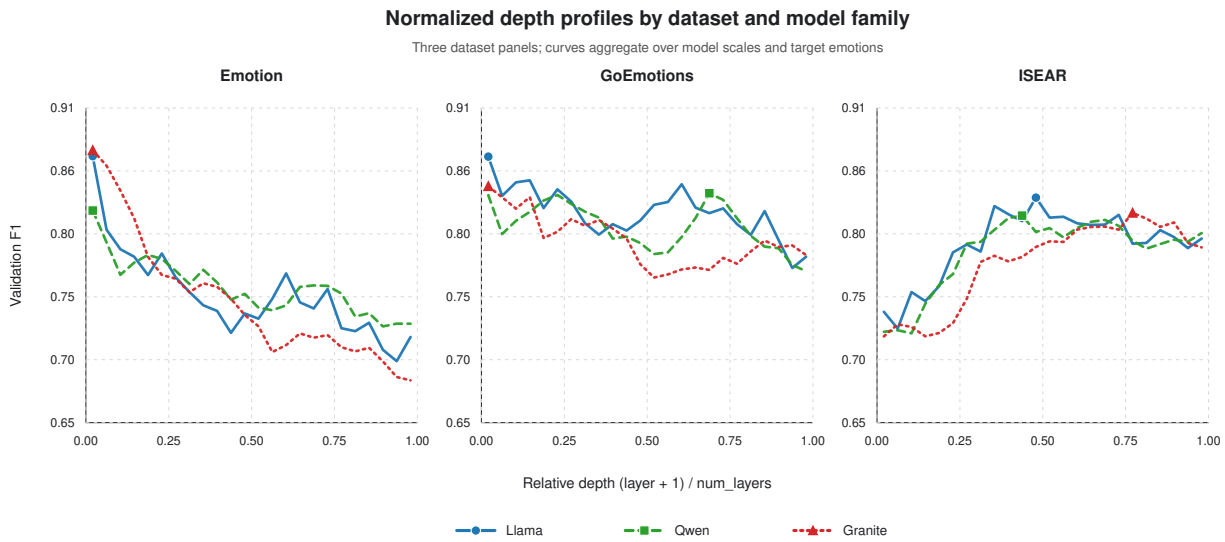

Figure 2: Layer-wise emotion decodability over normalized depth. Each panel corresponds to a dataset, and each curve aggregates validation F1 over model scales and target emotions within a model family. Across families, Emotion peaks early, GoEmotions is broader and intermediate, and ISEAR peaks later.

offline intervention that scales the cached frozen-feature representation. The online experiment provides the stronger causal evidence because the perturbation can propagate through later layers; the offline experiment checks whether the same bands matter in the static representation used by the probe.

5.1 Setup

For each model, dataset, target emotion, and band width $w \in \{1, 2, 3, 4\}$, we take the validation-selected emotion band from the layer-wise analysis. We compare it against a same-width *transfer* band selected for the same emotion on the other dataset, an *unrelated* non-overlapping band, and an unmodified baseline. During evaluation, a multiplicative hook scales the hidden states in the band by $\alpha \in \{0.0, 0.5, 1.5\}$; $\alpha = 0$ removes the band, $\alpha = 0.5$ attenuates it, and $\alpha = 1.5$ amplifies it.

All intervention claims use test-set predictions. We store per-example predictions for each condition and compute paired correctness contrasts within matched settings. Aggregate claims are summarized over settings with bootstrap confidence intervals, sign tests, Wilcoxon signed-rank tests, and BH-FDR correction within each comparison family.

5.2 Online intervention: main effect and specificity

Across all 768 forward settings, scaling the emotion-selected band causes a large accuracy drop relative to baseline (-26.6 percentage points, $q < 0.001$). The unrelated-band control is also disruptive (-20.5 points), showing that mid-forward perturbations are not harmless. The key comparison is therefore specificity: selected bands are more disruptive than unrelated same-width bands by -6.1 points overall ($q < 0.001$). At the strongest intervention, $\alpha = 0$, this specificity reaches -8.2, -13.4, -7.4, and -10.8 points for widths 1–4, respectively (Table 2). Thus, the layers identified by validation probing are not merely decodable; perturbing them affects the downstream readout more than perturbing arbitrary regions of the same size.

	W1	W2	W3	W4
$\alpha = 0.0$	-8.2^{*}	-13.4^{***}	-7.4	-10.8^{**}
$\alpha = 0.5$	-8.4^{*}	-6.2	-6.6	-9.1
$\alpha = 1.5$	-2.1	-0.5	-2.0	$+1.7$

Table 2: Forward specificity: selected-band minus unrelated-band accuracy delta, in percentage points. Negative values mean selected bands are more disruptive. Stars use BH-FDR-corrected sign tests: $^{*}q < 0.05$, $^{**}q < 0.01$, $^{***}q < 0.001$.

Cross-dataset transfer. Transfer bands, selected for the same emotion on the other dataset, produce nearly the same disruption as own-dataset bands. The own-vs-transfer difference is small and not reliable overall (-1.2 points, n.s.), while transfer bands are still more disruptive than unrelated bands (-4.9 points, $q < 0.001$). This supports partial cross-dataset sharing: an ISEAR-selected fear band, for example, can remain causally rele-

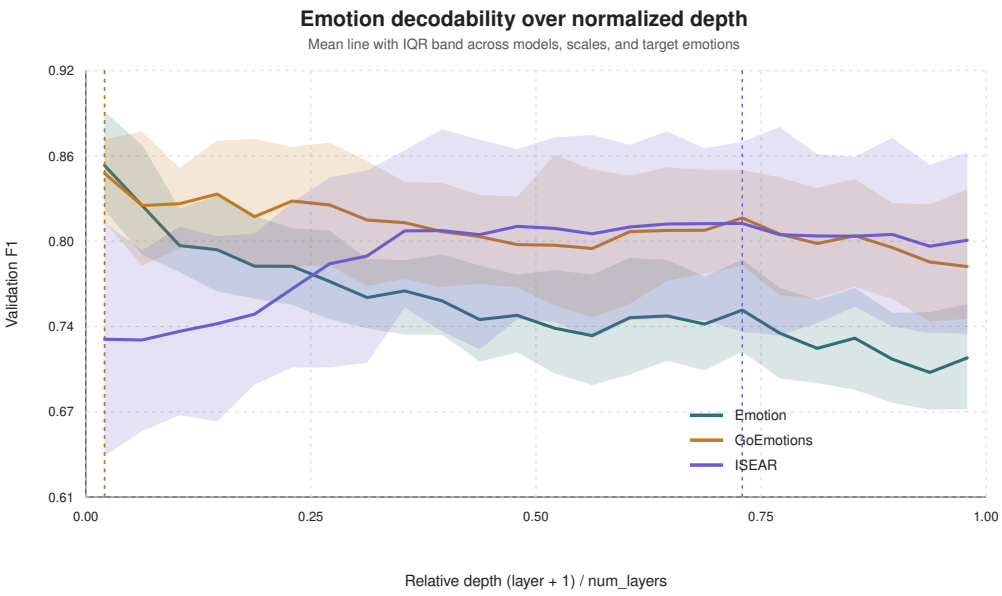

Figure 3: Dataset-level mean validation F1 over normalized depth, with IQR shading across models, scales, and target emotions. Emotion is early and sharp; GoEmotions is broader, more diffuse; ISEAR peaks substantially later.

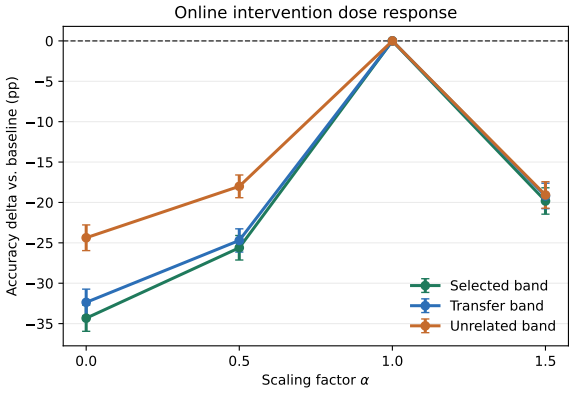

Figure 4: Online intervention dose response. Selected and transferred emotion bands are more disruptive than unrelated same-width bands, especially when $\alpha < 1$. Values are test accuracy deltas relative to the unmodified baseline. The $\alpha = 1$ point denotes the unmodified baseline and is shown only for reference.

vant during an Emotion forward pass, although the match is not exact.

Moderators. The specificity gap is strongest for Llama (-7.8 points) and Qwen (-9.3 points), both significant after correction, but not for Granite ($+1.2$ points, n.s.). Dataset also matters. On Emotion, selected bands are reliably more disruptive than unrelated bands (-7.7 points, $q < 0.001$), with significant specificity at every width. On ISEAR, the same contrast is smaller (-4.5 points) and less stable. In contrast, transfer-vs-unrelated is strongest on ISEAR (-12.2 points), suggesting that deeper narrative-style cues may be more compatible with transferred bands.

5.3 Offline feature scaling

The offline intervention applies the same scaling operation after feature extraction, directly to the cached representation read by the probe. It therefore tests representational usefulness, not causal computation inside the transformer. Offline scaling reproduces the main disruption effect but with weaker specificity: own-band scaling reduces accuracy by -12.2 points relative to baseline ($q < 0.001$), while unrelated and low-sensitivity bands also reduce accuracy by -11.3 and -10.8 points. The selected-vs-unrelated contrast is only -0.8 points and is not reliable. This contrast between online and offline results suggests that the causal evidence comes primarily from intervening during the forward computation, not merely from editing the final feature vector.

Cross-emotion transfer. The offline setting additionally lets us test cross-emotion transfer: scaling a band selected for a different emotion on the same dataset. At $\alpha = 0$, cross-emotion transfer is almost as disruptive as the own-emotion band (-15.2 vs. -15.9 points). Fear is the most vulnerable target (-16.8), joy the most robust (-13.1), and sadness is the most disruptive source (-16.5 on average). These patterns are exploratory, but they suggest that selected bands encode shared affective structure rather than completely emotion-specific modules.

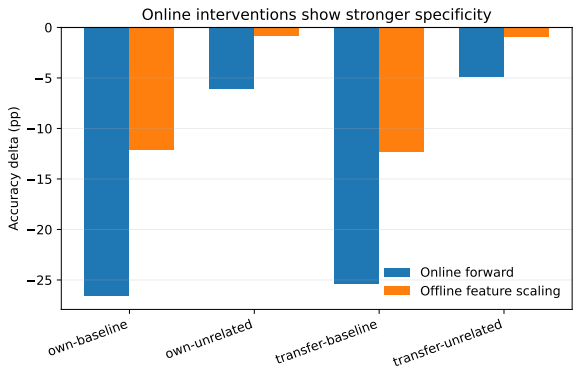

Figure 5: Online vs. offline intervention effects. Both settings show large disruption relative to baseline, but selected-vs-unrelated specificity is much clearer online. In the offline setting, transfer pools the available transfer conditions; own-vs-baseline and own-vs-unrelated are directly comparable across settings.

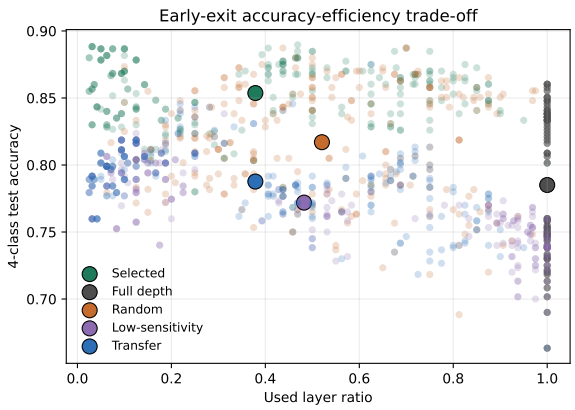

Figure 6: Early-exit accuracy-efficiency trade-off. Each small point is a model–dataset–emotion–width setting; large points show condition means. Selected bands achieve higher accuracy while using fewer layers than full-depth representations.

6 Early Exit from Emotion-Sensitive Bands

The intervention results show that emotion-selected bands are functionally important for the probe-mediated readout. We next ask the constructive question: can these bands support downstream classification without running the full transformer?

Setup. For each model and dataset, we train a 4-class multinomial logistic regression head on `concat(mean,max,min)` pooled hidden states. For early exit, the transformer is stopped at the last layer in the selected band. We compare selected bands against full-depth final-layer representations, random same-width bands, low-sensitivity bands, and cross-dataset transfer bands. As above, all aggregate significance tests use matched test-set predictions with BH-FDR correction.

Selected bands outperform full depth and controls. Selected early exits outperform full-depth representations by $+6.9$ accuracy points overall ($q < 0.001$). They also outperform random bands ($+3.7$), low-sensitivity bands ($+8.2$), and transferred bands ($+6.6$), all with $q < 0.001$. These comparisons rule out the trivial explanation that any intermediate representation works equally well: the validation-selected emotion bands are consistently better readout points for the lightweight classifier.

Bandwidth. Early-exit gains are not monotonic in bandwidth. Width 1 gives the largest selected-vs-full-depth gain ($+7.8$ points), width 3 is the weakest ($+5.9$), and width 4 partially recovers ($+6.5$). This matches the layer-wise picture: when emotion cues are concentrated, a narrow band can capture the useful signal without adding off-peak features; when cues are more distributed, wider bands add capacity but with diminishing returns.

Early-exit transfer. Early exit exposes two notions of transfer. *Layer transfer* keeps the classifier trained on the target dataset but exits at a band selected on the other dataset. These transferred bands roughly match full depth overall ($+0.3$ points, n.s.) but remain below target-selected exits by 6.6 points. The effect is asymmetric: transfer exits help on Emotion relative to full depth ($+3.7$ points) but hurt on ISEAR (-3.2 points). *Source-train transfer* also changes the classifier training data. In that setting, source-selected bands slightly outperform target-selected bands under the source-trained head ($+1.2$ points, $q < 0.01$), suggesting that exit quality depends both on where emotion is encoded and on which representation the classifier was trained to read.

Frozen-encoder comparison. We also compare selected early-exit LLM representations with frozen RoBERTa-base and DeBERTa-v3-base representations under the same logistic-regression protocol. We keep all encoders frozen because our goal is to evaluate where emotion-related information is already available in pretrained representations, rather than how much performance can be obtained after task-specific parameter updates. Fully fine-tuning an encoder can reshape its internal representations, making it less directly comparable to the layer-localization and early-exit set-

ting studied here. Selected LLM exits outperform frozen RoBERTa by $+15.8$ points and frozen DeBERTa by $+21.5$ points overall (Table 14). This is a frozen-representation comparison, not a fine-tuned SOTA claim, but it shows that the selected LLM bands provide a strong task representation under a lightweight head.

Moderators. The dataset selection is the strongest moderator. Selected-vs-full-depth gain is +11.3 points on Emotion but only +2.5 on ISEAR, consistent with Emotion's shallow and concentrated layer-wise profile. Model family also matters: Granite shows the largest gain (+8.8), followed by Llama (+6.9) and Qwen (+5.5).

7 Discussion

Overall, emotion-sensitive information is localizable enough for intervention and early readout, but its layer-wise organization varies with emotion expression and readout setting.

Emotion depth reflects how affect is expressed. The clearest pattern is that the best emotion layers shift with dataset source and style. This supports a more nuanced view than the claim that affective information simply lives in early layers. Early layers are sufficient for CARER-like short posts, where emotion is often marked by lexical or conventional cues. In contrast, ISEAR narratives require deeper representations, plausibly because the emotion must be inferred from events, appraisals, and causal context. GoEmotions falls between these cases, with a broader and less sharply localized profile. Thus, the depth of emotion decodability should be treated as an empirical property of the input distribution, not only as a property of the model.

Intervention and readout reveal different notions of usefulness. The online intervention results show that probe-selected bands are not only diagnostic: perturbing them disrupts downstream prediction more than matched unrelated bands. However, the offline scaling results are less specific, which suggests that the causal role of a band depends on its participation in the forward computation, not only on the final cached feature vector. Early-exit results add a constructive perspective: some intermediate bands are better readout points than the final layer under a lightweight classifier. Together, these findings separate three questions that are often conflated: where emotion is decodable, where it is causally involved, and where it is easiest to read out.

Transfer is shared but readout-dependent. Transfer results show that emotion-sensitive regions are not strictly dataset-specific or label-specific. Bands selected on one dataset can remain disruptive on another, and different-emotion bands can affect the same target. At the same time, early-exit transfer is weaker, indicating that a useful causal region is not necessarily an optimal readout point for a classifier trained in a different setting. This distinction helps explain why transfer appears stronger in intervention than in early-exit classification: the former tests whether a region matters during computation, while the latter tests whether a trained classifier can use that region effectively.

Implications. These results suggest that affective information in LLMs is organized at a band level rather than at a single universal layer. Future analyses of emotion in LLMs should therefore report not only whether emotion is decodable, but also the dataset source, model family, layer-selection criterion, and readout mechanism under which the claim holds.

8 Conclusion

This paper examined whether emotion in decoder-only LLMs has a fixed layer-wise location, or whether its organization varies with the kind of text being processed. Our results support the latter. Emotion information is not uniformly shallow, nor does it occupy a single universal layer across models and datasets. Instead, the depth and breadth of emotion-sensitive bands vary systematically with dataset style and model family.

These bands are not merely correlational artifacts of a linear probe. They can be causally perturbed through forward intervention and reused as compact early-exit representations, while their cross-dataset and cross-emotion transfer suggests shared affect-sensitive regions rather than isolated label-specific modules.

Taken together, our findings recast emotion representation as a depth-sensitive and context-dependent property of pretrained LLMs. This view helps explain why prior layer-wise analyses can disagree across datasets, and suggests that future work on affective behavior in LLMs should treat where emotion is represented as part of the empirical question, not as a fixed architectural fact.

Limitations

Our analysis combines layer-wise probing, causal intervention, and early-exit experiments on eight open-weight decoder-only LLMs and three English emotion corpora under a fixed pooling and probe configuration. Several aspects of this design constrain the generality of our conclusions.

Dataset and language scope. We restrict our analysis to three English-language corpora (Emotion/CARER, ISEAR, GoEmotions) and to the four-emotion intersection of their original label sets: fear, joy, anger, and sadness. Other emotions such as surprise, disgust, shame, guilt, and love are not analyzed here. All three sources are English and largely Western, so the dataset-depth ordering we observe may not generalize to other languages, cultural contexts, or finer-grained emotion taxonomies.

Model scope. We probe eight open-weight decoder-only LLMs from three families (Llama, Qwen, Granite) between 1B and 9B parameters. We do not evaluate closed-weight models, base non-instruction-tuned checkpoints, or much larger open models. The model-family and scale patterns should therefore be read as descriptive findings within this open-weight range, not as universal claims about all decoder-only LLMs.

Methodological choices. All main experiments keep model parameters frozen and use deterministic pooling with a linear probe. This design is intentional: our goal is to study where emotion-related information is already available in pretrained representations, rather than how much task-specific performance can be obtained after updating model parameters. Fully fine-tuning an encoder or decoder could substantially reshape internal representations, making it less directly comparable to the layer-localization and early-exit setting studied here. A learnable attention-pooling baseline can reach higher absolute F1 in pilot runs, but it introduces additional training instability and capacity differences, making representation comparisons harder to interpret. Our interventions also operate at a coarse contiguous layer-band granularity. Finer-grained approaches such as activation patching, direction-level editing, nullspace projection, or neuron-level interventions could localize affective computation more precisely.

Finding-level caveats. Forward specificity is significant for Llama and Qwen but not Granite. Offline feature scaling reproduces the main disruption direction but does not show reliable selected-vs-unrelated specificity, consistent with our view that online intervention is the stronger causal diagnostic. Cross-emotion transfer (Table 10) suggests shared affective bands rather than strict per-emotion modules. The early-exit width effect is non-monotonic, and our explanation in terms of selectivity versus capacity remains exploratory. Source-train transfer is consistent with a classifier–band alignment hypothesis, but does not prove it.

Statistical and reproducibility caveats. We compute per-setting paired correctness contrasts from per-example test predictions. Aggregate claims are summarized across settings using bootstrap confidence intervals, sign tests, Wilcoxon signed-rank tests, and BH-FDR correction within each comparison family. Per-setting McNemar tests are available in the supplementary CSVs. We do not apply more conservative family-wise correction across every comparison in the paper. Finally, ISEAR is loaded through a community-uploaded HuggingFace dataset, so exact ISEAR-specific reproducibility depends on that source remaining unchanged.

Potential risks. This work is a preliminary representation-level analysis using publicly available emotion datasets, and does not propose or deploy an emotion-recognition system. Nevertheless, methods for localizing emotion-sensitive representations could be misused for affective profiling, large-scale emotion inference, or emotionally manipulative applications. We therefore present our findings as tools for understanding model behavior, not as recommendations for deploying emotion inference in sensitive or high-stakes settings.

References

Guillaume Alain and Yoshua Bengio. 2017. Understanding intermediate layers using linear classifier probes.

Alexandra Balahur, Jesús M. Hermida, and Andrés Montoyo. 2011. Detecting implicit expressions of sentiment in text based on commonsense knowledge. In *Proceedings of the 2nd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis (WASSA 2.011)*, pages 53–60, Portland, Oregon. Association for Computational Linguistics.

Yonatan Belinkov. 2022. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219.

Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. 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. Association for Computational Linguistics.

Dario Di Palma, Alessandro De Bellis, Giovanni Servedio, Vito Walter Anelli, Fedelucio Narducci, and Tommaso Di Noia. 2025. LLaMAs have feelings too: Unveiling sentiment and emotion representations in LLaMA models through probing. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6124–6142, Vienna, Austria. Association for Computational Linguistics.

Eli Dresner and Susan C. Herring. 2010. Functions of the nonverbal in cmc: Emoticons and illocutionary force. *Communication Theory*, 20(3):249–268.

Paul Ekman. 1992. An argument for basic emotions. *Cognition and Emotion*, 6(3-4):169–200.

Angela Fan, Edouard Grave, and Armand Joulin. 2020. Reducing transformer depth on demand with structured dropout. In *International Conference on Learning Representations*.

Mor Geva, Jasmijn Bastings, Katja Filippova, and Amir Globerson. 2023. Dissecting recall of factual associations in auto-regressive language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12216–12235, Singapore. Association for Computational Linguistics.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Wes Gurnee and Max Tegmark. 2024. Language models represent space and time. In *The Twelfth International Conference on Learning Representations*.

John Hewitt and Percy Liang. 2019. Designing and interpreting probes with control tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2733–2743, Hong Kong, China. Association for Computational Linguistics.

Jan Hofmann, Enrica Troiano, Kai Sassenberg, and Roman Klinger. 2020. Appraisal theories for emotion classification in text. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 125–138, Barcelona, Spain (Online). International Committee on Computational Linguistics.

IBM Research. 2026. Introducing the ibm granite 4.1 family of models. `https://research.ibm.com/blog/granite-4-1-ai-foundation-models`.

Dongjin Kang, Sunghwan Kim, Taeyoon Kwon, Seungjun Moon, Hyunsouk Cho, Youngjae Yu, Dongha Lee, and Jinyoung Yeo. 2024. Can large language models be good emotional supporter? mitigating preference bias on emotional support conversation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15232–15261, Bangkok, Thailand. Association for Computational Linguistics.

Samuel Marks and Max Tegmark. 2024. The geometry of truth: Emergent linear structure in large language model representations of true/false datasets. In *First Conference on Language Modeling*.

Thomas McGrath, Matthew Rahtz, Janos Kramar, Vladimir Mikulik, and Shane Legg. 2023. The hydra effect: Emergent self-repair in language model computations. *Preprint*, arXiv:2307.15771.

Saif Mohammad and Felipe Bravo-Marquez. 2017. Emotion intensities in tweets. In *Proceedings of the 6th Joint Conference on Lexical and Computational Semantics (*SEM 2017)*, pages 65–77, Vancouver, Canada. Association for Computational Linguistics.

Saif M. Mohammad and Svetlana Kiritchenko. 2015. Using hashtags to capture fine emotion categories from tweets. *Comput. Intell.*, 31(2):301–326.

Robert Plutchik. 1980. A general psychoevolutionary theory of emotion. In Robert Plutchik and Henry Kellerman, editors, *Emotion: Theory, Research, and Experience*, volume 1, pages 3–33. Academic Press.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. 2020. A primer in BERTology: What we know about how BERT works. *Transactions of the Association for Computational Linguistics*, 8:842–866.

Sahand Sabour, Siyang Liu, Zheyuan Zhang, June Liu, Jinfeng Zhou, Alvionna Sunaryo, Tatia Lee, Rada Mihalcea, and Minlie Huang. 2024. 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)*, pages 5986–6004, Bangkok, Thailand. Association for Computational Linguistics.

Hassan Sajjad, Fahim Dalvi, Nadir Durrani, and Preslav Nakov. 2023. On the effect of dropping layers of pre-trained transformer models. *Comput. Speech Lang.*, 77(C).

Elvis Saravia, Hsien-Chi Toby Liu, Yen-Hao Huang, Junlin Wu, and Yi-Shin Chen. 2018. CARER: Contextualized affect representations for emotion recognition. In *Proceedings of the 2018 Conference on*

Empirical Methods in Natural Language Processing, pages 3687–3697, Brussels, Belgium. Association for Computational Linguistics.

K. R. Scherer and H. G. Wallbott. 1994a. Evidence for universality and cultural variation of differential emotion response patterning. *Journal of Personality and Social Psychology*, 66(2):310–328.

Klaus R. Scherer and Harald G. Wallbott. 1994b. Evidence for universality and cultural variation of differential emotion response patterning. *Journal of Personality and Social Psychology*, 66(2):310–328.

Tal Schuster, Adam Fisch, Jai Gupta, Mostafa Dehghani, Dara Bahri, Vinh Q. Tran, Yi Tay, and Donald Metzler. 2022. Confident adaptive language modeling. In *Advances in Neural Information Processing Systems*.

Roy Schwartz, Gabriel Stanovsky, Swabha Swayamdipta, Jesse Dodge, and Noah A. Smith. 2020. The right tool for the job: Matching model and instance complexities. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6640–6651, Online. Association for Computational Linguistics.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019. BERT rediscovers the classical NLP pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4593–4601, Florence, Italy. Association for Computational Linguistics.

Enrica Troiano, Laura Oberländer, and Roman Klinger. 2023. Dimensional modeling of emotions in text with appraisal theories: Corpus creation, annotation reliability, and prediction. *Computational Linguistics*, 49(1):1–72.

Eric Wallace, Yizhong Wang, Sujian Li, Sameer Singh, and Matt Gardner. 2019. Do NLP models know numbers? probing numeracy in embeddings. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5307–5315, Hong Kong, China. Association for Computational Linguistics.

Xuena Wang, Xueting Li, Zi Yin, Yue Wu, and Jia Liu. 2023. Emotional intelligence of large language models. *Journal of Pacific Rim Psychology*, 17:18344909231213958.

Ji Xin, Raphael Tang, Jaejun Lee, Yaoliang Yu, and Jimmy Lin. 2020. DeeBERT: Dynamic early exiting for accelerating BERT inference. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2246–2251, Online. Association for Computational Linguistics.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

A Per-emotion Dataset Statistics

This appendix reports the per-emotion sample distribution underlying the aggregate dataset sizes in Table 1.

B Pooling and Probe Selection Pilot

This appendix details the two-step greedy selection protocol used to fix the pooling function and probe classifier. All reported numbers are validation F1 at the best-performing layer per (emotion, pooling, probe) cell.

Step 1: Pooling selection. We fix the probe to a linear SVM on Llama-3.2-3B-Instruct and scan all layers of the Emotion validation split for each of the four target emotions, comparing last-token, mean, and `concat(mean, max, min)` pooling (Table 4). `concat` is the strongest non-parametric option on all four emotions, outperforming last-token by 24 to 35 F1 points and mean by 2 to 8 F1 points.

Max/min ablation. To verify that the `concat` advantage is not driven by the max or min component alone, we additionally compare `concat` against max and min pooling on fear and joy with the same Llama-3B linear SVM probe (Table 5). `concat` remains the strongest option in both cases.

Cross-family pooling sanity check. We replicate the pooling comparison on Qwen-3.5-4B for fear (one emotion only). The ranking is consistent with Llama-3B: `concat` (F1 0.822) is strongest, followed by mean (0.795) and last-token (0.536).

Step 2: Probe selection. With `concat(mean, max, min)` pooling fixed, we compare three lightweight probes on the same Llama-3B four-emotion validation setup: linear SVM, logistic regression, and a one-hidden-layer MLP (Table 6). Logistic regression is the strongest probe on fear and joy and competitive on anger and sadness (within 0.01 F1 of the best); we adopt it as the fixed analysis probe because it is also simpler and more stable than the MLP alternative.

Cross-family probe sanity check. We replicate the probe comparison on Qwen-3.5-4B for fear: linear SVM 0.843, logreg 0.847, MLP 0.832. Logistic regression is again the strongest probe.

Dataset	Split	Fear	Joy	Anger	Sadness	Total
Emotion (CARER)	Train	1,200	1,200	1,200	1,200	4,800
	Val	150	150	150	150	600
	Test	150	150	150	150	600
ISEAR	Train	823	816	831	764	3,234
	Val	102	102	103	95	402
	Test	104	102	105	97	408
GoEmotions	Train	321	604	644	578	2,147
	Val	68	129	138	124	459
	Test	70	130	138	125	463

Table 3: Per-emotion sample counts in the four-emotion splits of each dataset.

Pooling	Fear	Joy	Anger	Sadness	Mean
last-token	0.554	0.628	0.551	0.494	0.557
mean	0.829	0.876	0.794	0.743	0.811
concat	0.908	0.880	0.836	0.766	0.848

Table 4: Step-1 pooling pilot on Llama-3.2-3B-Instruct, Emotion validation split, linear SVM probe. Best-layer F1 per emotion.

Pooling	Fear	Joy
concat	0.908	0.880
max	0.878	0.818
min	0.843	0.836

Table 5: Max/min ablation on Llama-3.2-3B-Instruct, Emotion validation split, linear SVM probe.

Verification: learned attention pooling. As a verification of our restriction to non-parametric pooling, we evaluate a single-head learnable attention pooling layer followed by a linear classification head, trained end-to-end with cross-entropy on the Llama-3.2-3B-Instruct Emotion validation split. We use the same random seed (42) as our main experiments, a fixed learning rate of $\eta = 0.003$, and sweep over weight decay $\{0, 10^{-4}, 10^{-3}\}$ at each of six representative layers ($\{0, 3, 4, 20, 23, 26\}$), yielding 72 runs in total.

Table 7 reports F1 per (layer, emotion) averaged across the three weight decay values. At layer 0, attention pooling reaches strong F1 on all four emotions, but training collapses in 10 of the 24 (layer, emotion) cells at this seed: joy fails from layer 3 onward, fear from layer 20, and most emotions are reduced to $F1 = 0$ by layer 23. The mean F1 across the 24 cells is 0.51, well below the concat baseline mean of 0.85.

For completeness, when we cherry-pick the best (layer, weight decay) per emotion at this seed, the best-case mean F1 reaches 0.96 (fear 0.957, joy 0.963, anger 0.946, sadness 0.961), exceeding concat on each emotion. Achieving this best-case in practice would require either layer-by-layer collapse screening or multi-seed retries to find a configuration that converges, neither of which is feasible at the scale of our main experiments (8 models $\times$ 3 datasets $\times$ all transformer layers). In contrast, `concat(mean, max, min)` pooling is fully deterministic and applies uniformly without retuning. This per-layer instability also reflects the methodological cost of introducing learned parameters in the pooling stage: per-layer F1 measures not only the representation but also whether the attention parameters successfully trained at that layer.

Probe	Fear	Joy	Anger	Sadness	Mean
linear SVM	0.885	0.858	0.833	0.777	0.838
logreg	0.903	0.900	0.837	0.800	0.860
MLP	0.893	0.858	0.845	0.812	0.852

Table 6: Step-2 probe pilot on Llama-3.2-3B-Instruct, Emotion validation split, `concat(mean, max, min)` pooling. Best-layer F1 per emotion.

Layer	Fear	Joy	Anger	Sadness
0	0.95	0.96	0.93	0.93
3	0.94	0.00	0.94	0.96
4	0.59	0.00	0.94	0.94
20	0.00	0.00	0.94	0.92
23	0.00	0.00	0.92	0.00
26	0.00	0.00	0.40	0.00

Table 7: Per-(layer, emotion) F1 of attention pooling at $\eta = 0.003$, seed 42, averaged across three weight decay values. Zero entries indicate training collapse.

C Additional Layer-Wise Analyses

This appendix provides supplementary analyses for Section 4. All results are computed from validation-only layer-wise probing outputs. The main paper reports the condensed dataset- and family-level patterns; here we include the full one-vs-rest profiles, the four-class consistency check, cross-emotion correlations, and additional descriptive summaries.

Full normalized-depth profiles. Figure 7 shows the full 3×3 family-by-dataset normalized-depth view. This is the expanded version of Figure 2. The same dataset-level ordering remains visible within

model families: Emotion peaks early, GoEmotions is broader and intermediate, and ISEAR peaks later.

Emotion-specific layer-wise profiles. Figures 8, 9, and 10 show the normalized-depth profiles separately for each target emotion. These figures provide the detailed curves underlying the dataset-level summaries in the main text.

OvR versus four-class consistency. Figure 11 compares the best relative depth from the four-class multinomial probe against the mean best relative depth from the four one-vs-rest probes. Across all 24 dataset–model pairs, the correlation is $r = 0.85$, and 87.5% of four-class best layers fall within the range spanned by the four one-vs-rest best layers. This supports using one-vs-rest probes for target-specific band selection.

Cross-emotion curve correlations. Figure 12 reports correlations between emotion-specific F1-vs-depth curves. Correlations are positive across emotion pairs, suggesting that different emotions partly share layer-wise decodability structure. These correlations are descriptive and should not be interpreted as evidence that emotions share the same causal substrate.

D Forward Intervention Significance Details

The forward master CSV contains 768 matched test settings (8 models $\times$ 2 datasets $\times$ 4 emotions $\times$ 4 widths $\times$ 3 alpha values). For each comparison, we compute paired correctness contrasts from per-example predictions, then summarize across settings with bootstrap confidence intervals, sign tests, Wilcoxon signed-rank tests, and BH-FDR correction.

Comparison	n	Mean Δacc
Own vs. baseline	768	-26.6^{***}
Transfer vs. baseline	768	-25.4^{***}
Unrelated vs. baseline	768	-20.5^{***}
Own vs. transfer	768	-1.2
Own vs. unrelated	768	-6.1^{***}
Transfer vs. unrelated	768	-4.9^{***}

Table 8: Forward intervention aggregate effects in percentage points. Stars indicate BH-FDR-corrected significance.

E Offline Feature-Scaling Significance Details

The offline master CSV mirrors the forward setup but applies scaling to cached representations. It contains own/control settings plus cross-dataset and cross-emotion transfer settings. Because the transformer is not re-run after the edit, this analysis supports representational usefulness but is weaker causal evidence than the online intervention.

Comparison	n	Mean Δacc
Own vs. baseline	768	-12.2^{***}
Transfer vs. baseline	3072	-12.3^{***}
Unrelated vs. baseline	768	-11.3^{***}
Low-sensitivity vs. baseline	768	-10.8^{***}
Own vs. unrelated	768	-0.8
Own vs. low-sensitivity	768	-1.3
Transfer vs. own	3072	-0.1
Transfer vs. unrelated	3072	-1.0

Table 9: Offline feature-scaling aggregate effects in percentage points. Main disruption effects are reliable, but selected-band specificity is weak.

Target / Source	Anger	Fear	Joy	Sad.	Mean
Anger	–	-14.9	-14.6	-16.6	-15.4
Fear	-17.0	–	-16.7	-16.8	-16.8
Joy	-11.2	-12.0	–	-16.1	-13.1
Sad.	-14.3	-16.4	-15.7	–	-15.5
Mean	-14.2	-14.5	-15.7	-16.5	

Table 10: Offline cross-emotion transfer at $\alpha = 0$, in accuracy points. Row is the target probe; column is the emotion used to select the perturbed source band.

F Early Exit Significance Details

The early-exit master CSV combines target-trained controls, cross-dataset layer-transfer runs, and source-trained transfer runs. All reported tests use matched test-set predictions and BH-FDR correction within comparison families.

Comparison	n	Mean Δacc
Selected vs. full depth	256	$+6.9^{***}$
Selected vs. random	256	$+3.7^{***}$
Selected vs. low-sensitivity	256	$+8.2^{***}$
Selected vs. transfer	256	$+6.6^{***}$
Transfer vs. full depth	256	$+0.3$

Table 11: Early-exit aggregate effects in percentage points. Positive values mean the first condition is more accurate.

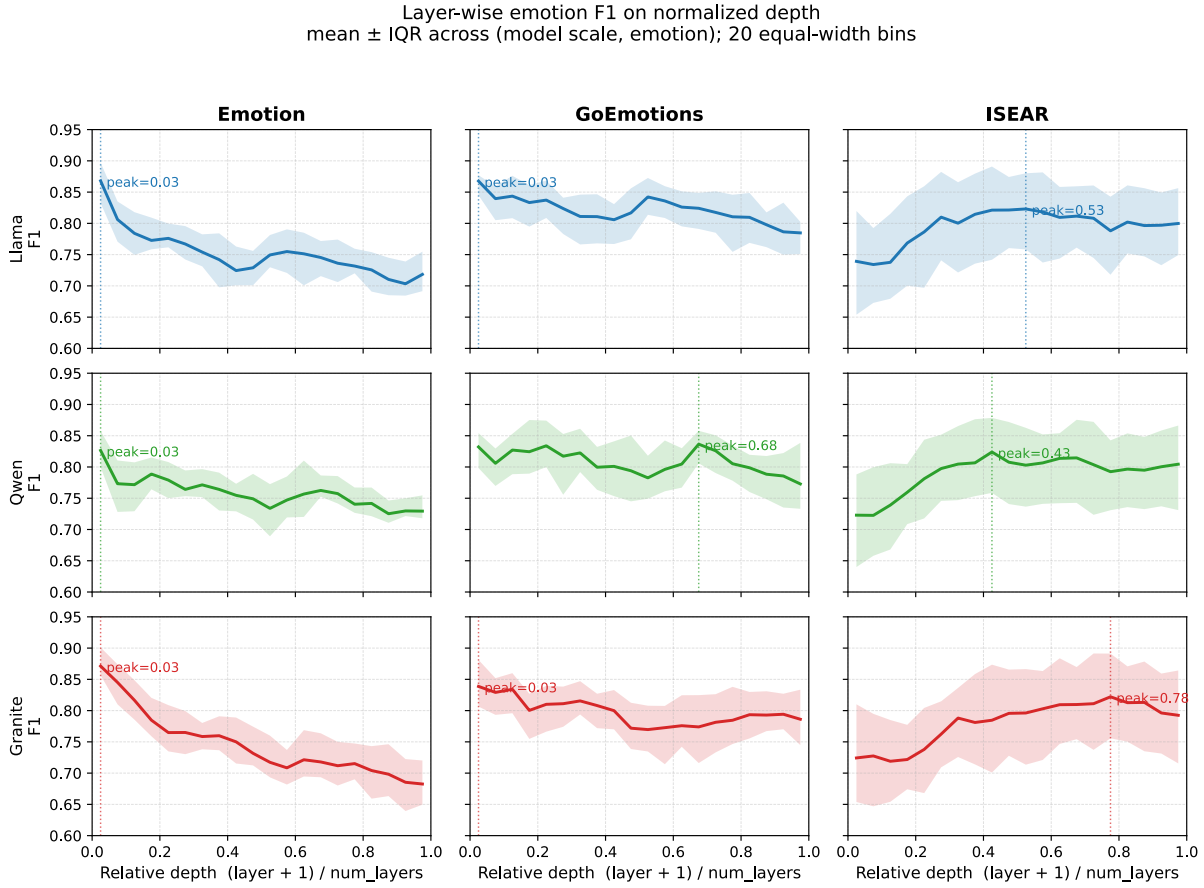

Figure 7: Full family-by-dataset normalized-depth profiles. Each panel aggregates validation F1 over model scales and target emotions within one model family and dataset.

Stratum	n	Selected vs. full depth
Emotion	128	$+11.3^{***}$
ISEAR	128	$+2.5^{***}$
Granite	64	$+8.8^{***}$
Llama	96	$+6.9^{***}$
Qwen	96	$+5.5^{***}$

Table 12: Early-exit selected-vs-full-depth gain by dataset and model family, in accuracy points.

Width	Selected vs. full depth	Selected vs. random
1	$+7.8^{***}$	$+4.5^{***}$
2	$+7.2^{***}$	$+3.8^{***}$
3	$+5.9^{***}$	$+2.7^{***}$
4	$+6.5^{***}$	$+3.6^{***}$

Table 13: Early-exit gains by band width, in accuracy points.

Baseline	n	Selected LLM exit gain
Frozen RoBERTa-base	256	$+15.8^{***}$
Frozen DeBERTa-v3-base	256	$+21.5^{***}$

Table 14: Frozen encoder baseline comparison under the same logistic-regression readout.

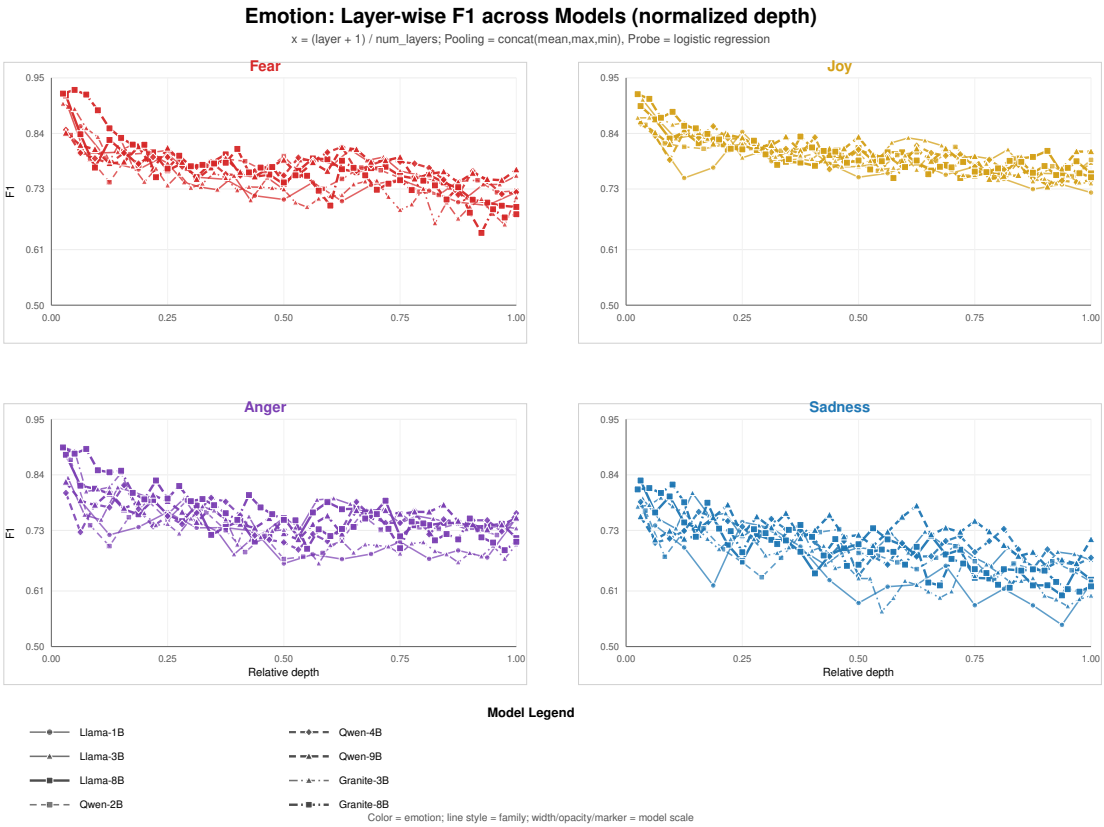

Figure 8: Emotion dataset: emotion-specific layer-wise validation F1 over normalized depth.

Figure 9: GoEmotions: emotion-specific layer-wise validation F1 over normalized depth.

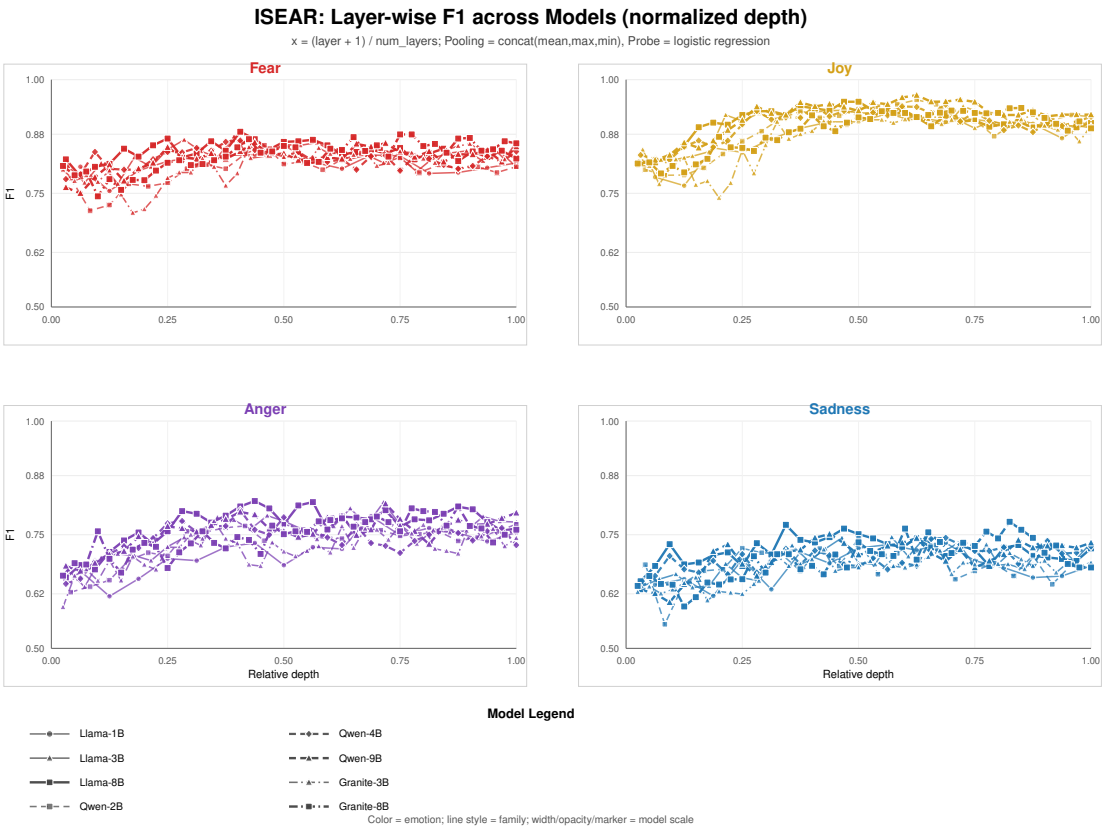

Figure 10: ISEAR: emotion-specific layer-wise validation F1 over normalized depth.

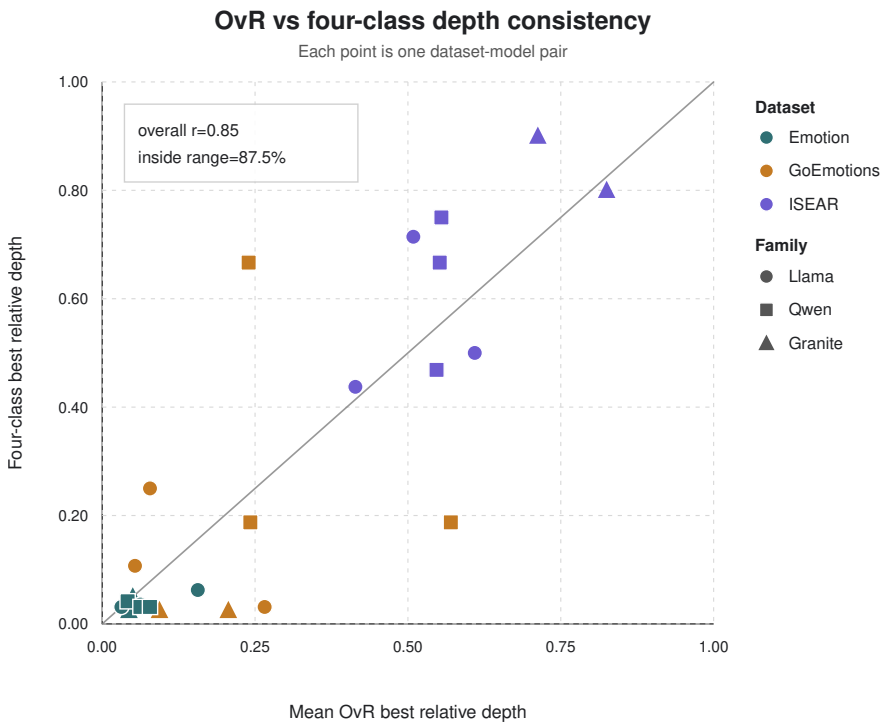

Figure 11: Consistency between one-vs-rest and four-class layer-wise probes. Each point is a dataset–model pair.

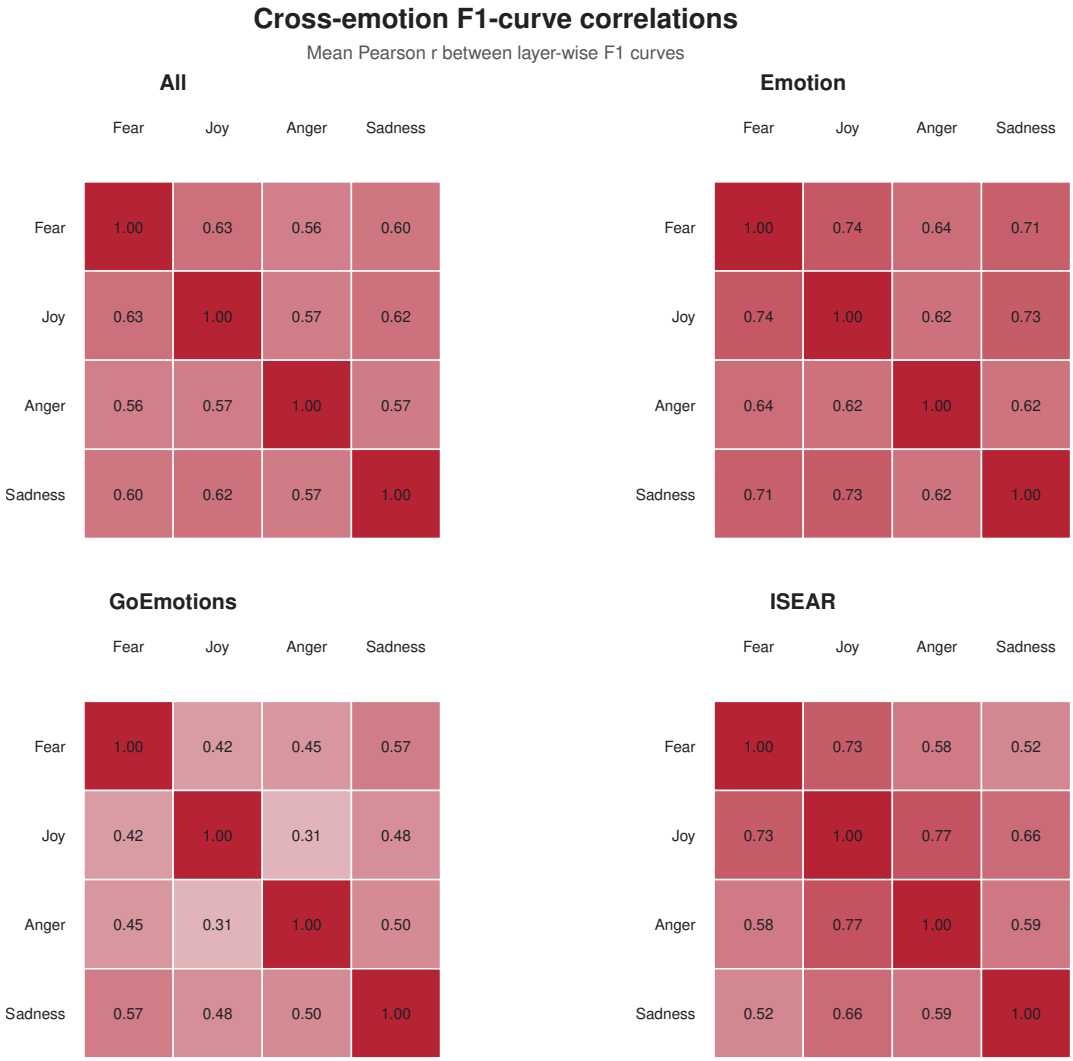

Figure 12: Cross-emotion correlations between layer-wise F1 curves, averaged across models.

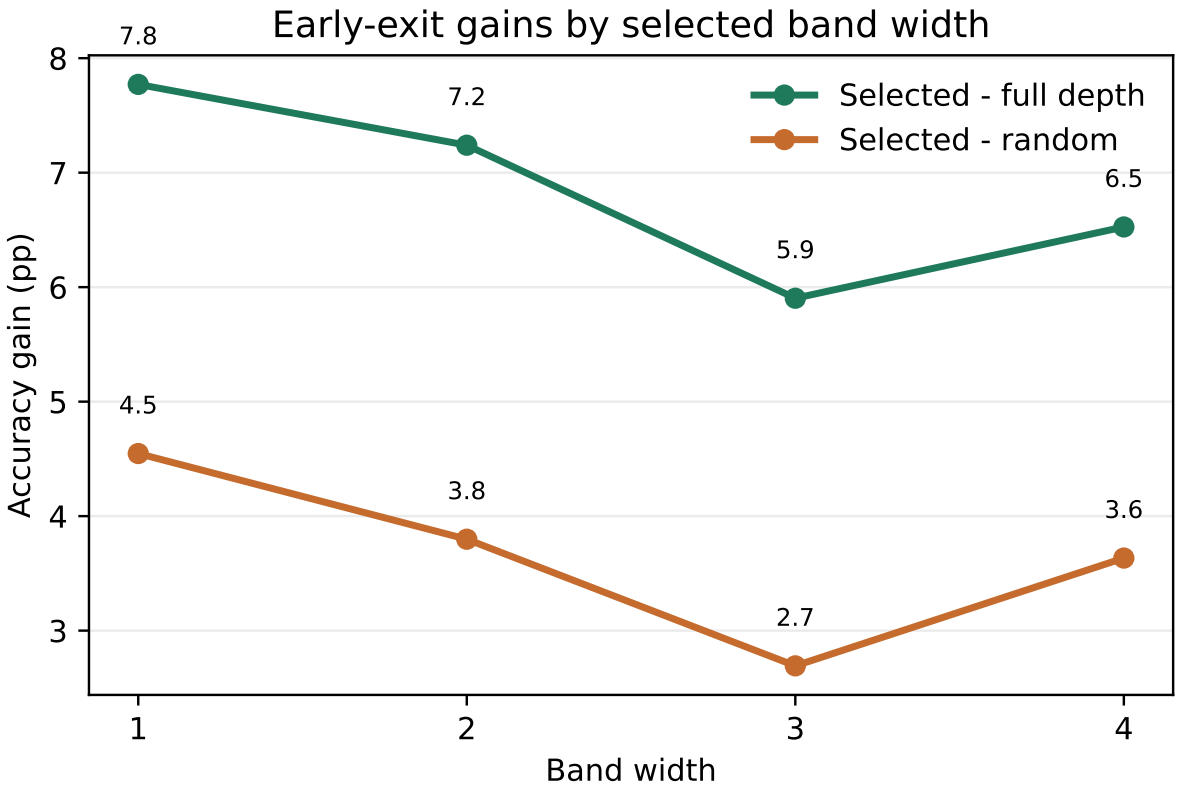

Figure 13: Early-exit gains by band width. Selected bands outperform both full-depth and random controls at every width, with the largest gain at width 1.