Negative Before Positive: Asymmetric Valence Processing in Large Language Models

Anonymous Authors[1]

Abstract

Mechanistic interpretability has revealed how concepts are encoded in large language models (LLMs) but emotional content remains poorly understood at the mechanistic level. We study whether LLMs process emotional valence through dedicated internal structure or through surface token matching. Using activation patching and steering on open-source LLMs, we find that negative and positive valence are processed at different network depths. Negative outcomes localize to early layers while positive outcomes peak at mid-to-late layers. Holding topic fixed while flipping valence produces sign-opposite responses, ruling out topic detection. Steering with the good-news direction at the identified layers shifts neutral prompts toward positive valence, showing these layers encode valence as a manipulable direction. Emotional valence in LLMs is localized, causal and steerable, making it a concrete target for interpretability-based oversight.

1. Introduction

Understanding what LLMs represent internally has become a central goal of mechanistic interpretability (Elhage et al., 2021; Olsson et al., 2022). Prior work has shown that LLMs encode factual associations in specific layers (Meng et al., 2022), implement recognizable algorithms (Wang et al., 2022) and represent abstract concepts as linear directions in the residual stream (Tigges et al., 2023; Park et al., 2023). Emotional content, however, has been studied primarily through probing and behavioral analysis rather than causal intervention. This matters: if models have internal valence representations that causally shape their outputs, those representations are directly relevant for understanding and controlling model behavior.

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

Sofroniew et al. (2026) recently found that Claude Sonnet 4.5 contains internal representations of emotional concepts that linearly predict and causally influence behavior, which they term functional emotions. This is an important result yet it leaves three questions open. First, the evidence is primarily correlational. Probing and steering demonstrate that emotion-like features exist but do not isolate which specific layers are causally responsible. Second, the study focuses on a single closed frontier model. Third, it does not test whether the observed signal reflects genuine valence tracking or topic detection. A model that simply recognizes the word "rejected" could produce the same probing signal as one with a rich internal valence representation. Without causal localization, it remains unclear whether emotional representations are concentrated within specific processing stages or distributed diffusely across the network.

We address these gaps using activation patching (Vig et al., 2020; Meng et al., 2022) on three open-source LLMs together with a prompt design that explicitly controls for topic confounds. Our results show a clear depth-wise dissociation in valence processing: negative outcomes are causally localized to early layers while positive outcomes peak at mid-to-late layers. This pattern is consistent across all evaluated models. We further validate that the effect is genuinely valence-specific through a topic-controlled flip test. Using a shared-corrupted-baseline design, prompt pairs reverse sign significantly above chance when valence changes. Finally, causal steering experiments demonstrate that the identified representations behave as manipulable linear directions. Adding or subtracting them from the residual stream predictably alters responses to emotionally neutral inputs.

2. Related Work

Mechanistic interpretability. Activation patching has been widely used to study factual recall and causal circuits in transformers (Meng et al., 2022; Wang et al., 2022; Conmy et al., 2023). Lieberum et al. (2023) examine whether circuit analysis techniques scale to larger models, highlighting challenges in extending interpretability methods beyond small settings. The residual stream framework (Elhage et al., 2021) and logit lens (nostalgebraist, 2020) provide

theoretical grounding for layer-wise analysis of internal representations.

Emotion and sentiment in LLMs. Tigges et al. (2023) and Marks & Tegmark (2023) identify linearly decodable sentiment and truth directions in intermediate transformer layers using probing methods. Tak et al. (2025) provide a mechanistic analysis of emotion inference across model families, finding concentration of emotion-relevant computations in specific layers. Zhang & Zhong (2025) further separate affect reception and emotion categorization across depth, suggesting functional specialization across layers. Wang et al. (2025) study emotion circuits in LLMs by identifying neurons and attention heads associated with emotional expression and demonstrate controllable emotion generation through circuit-level interventions. Hofmann et al. (2024) show that transformer representations encode socially relevant attributes beyond surface text patterns.

Steering vectors. Steering vectors have been shown to correspond to directions in activation space that reliably control model behavior (Turner et al., 2024; Zou et al., 2023). They have been applied to identify and modify specific behaviors such as refusal (Arditi et al., 2024) and to study robustness across inputs (Panickssery et al., 2023).

Representation geometry. The linear representation hypothesis posits that semantic concepts are encoded as linear directions in activation space (Park et al., 2023; Elhage et al., 2022). Recent work supports the existence of monosemantic or highly interpretable features at scale (Bricken et al., 2023; Templeton et al., 2024), suggesting structured geometry in learned representations.

Our work connects these lines by studying whether valence exhibits layer-dependent causal structure and whether it can be expressed as a manipulable linear direction in residual stream space.

3. Methods

3.1. Methodological Background

Our analysis builds on two core interpretability techniques: activation patching and steering vectors. Activation patching (Vig et al., 2020; Meng et al., 2022) is a causal intervention method for identifying where in a network information relevant to a behavior is computed. Given a clean and a corrupted input that differ in a controlled way, we run the model on the corrupted input while replacing residual stream activations at a given layer with those from the clean run. If this restores the clean output behavior, that layer is considered causally important for the computation. We sweep all layers and measure effects using a logit gap metric defined in Section 3.4.

Steering vectors (Turner et al., 2024; Zou et al., 2023) are linear directions in activation space that can be added to residual stream activations at inference time to systematically shift model behavior. They are constructed from differences between activations under contrasting conditions and correspond to directions in representation space associated with latent features. We extract steering vectors from the residual stream at the most causally relevant layer identified by patching.

3.2. Models

We study three instruct-tuned models that span two architectural families and two scales: Llama-3.2-1B-Instruct (Meta, 2024), Qwen2.5-1.5B-Instruct, and Qwen2.5-3B-Instruct (Yang et al., 2024). The Llama model employs Multi-head attention (Cordonnier et al., 2020) while the Qwen models use Grouped-query attention (Ainslie et al., 2023). Llama-1B vs. Qwen-1.5B provides an approximate architecture comparison (MHA vs. GQA), while Qwen-1.5B vs. Qwen-3B provides a scale comparison within the same attention architecture. Since Llama-1B and Qwen-1.5B differ in both architecture and parameter count, the architecture comparison is only approximate.

MHA uses independent key-value projections per head. GQA shares key-value projections across groups of heads, changing how information is routed between layers. We study whether the layer dissociation finding holds across both variants. All models use a fixed system prompt: "You are a concise assistant. Respond in one or two sentences." All experiments use `float16` precision on a single GPU.

3.3. Prompt Pair Design

We construct 100 clean/corrupted prompt pairs per condition. Each pair consists of a clean prompt and a corrupted prompt that differ in exactly one dimension: emotional valence. We use two conditions:

- **Good news:** the clean prompt describes a positive outcome; the corrupted prompt describes the same situation neutrally with no emotional signal.
- **Negative control:** the clean prompt describes a negative outcome; the corrupted prompt is *identical* to the good-news corrupted baseline.

The shared corrupted baseline is the core design decision. It means both conditions are patched into the exact same starting point, so any difference in patch effect can only be attributed to the valence of the clean prompt and not to topic or phrasing. A model doing pure topic detection would produce similar patch effects from both clean runs since the topic is the same. A model with a genuine internal valence

representation would produce sign-opposite effects since the valence is opposite. Example pair:

> **Clean (good news):** "I just got accepted into my dream PhD program today."
>
> **Clean (negative):** "I just got rejected from my dream PhD program today."
>
> **Corrupted (shared):** "I just received an email about my PhD program today."

Pairs are drawn from three broad domains: academia, career and personal life. These domains were chosen to represent situations where emotional outcomes are common and clearly valenced while covering distinct vocabularies. If the layer dissociation held only in academia prompts it could reflect domain-specific vocabulary and not valence processing. We verify it holds across all three domains in Appendix B.

Pairs where the clean and corrupted prompts tokenize to different lengths are left-padded to equal length using the model's pad token (or EOS as fallback). This preserves right-alignment of content tokens so the final token position is always a real token.

3.4. Valence Metric

We measure a model's emotional lean by computing a logit gap between two sets of anchor tokens at the next-token distribution.

Anchor tokens. Positive anchors include *congratulations, congrats, happy, glad, wonderful, amazing, thrilled, proud, fantastic, excellent* while negative anchors include *okay, noted, fine, ordinary, received, sorry*. We include only tokens that tokenize as a single unit for a given model since multi-token words have split probability mass and produce unreliable logit comparisons. The number of valid anchors varies by model and is reported in each experiment.

Score definition.

$$\text{score} = \frac{1}{|\mathcal{P}|} \sum_{i \in \mathcal{P}} \log p_i - \frac{1}{|\mathcal{N}|} \sum_{j \in \mathcal{N}} \log p_j \qquad (1)$$

A positive score indicates the model's next-token distribution leans toward positive-valence tokens. A negative score indicates a lean toward negative-valence tokens. The score gap between clean and corrupted runs measures how strongly the emotional content of the clean prompt shifts the model's output distribution.

Metric validity. The metric is a proxy: it measures next-token probability mass on a fixed anchor set but does not directly measure an internal emotional state. The main risks are anchor sensitivity, where different token sets can produce different scores, and ceiling effects, where models that strongly favor a specific token can artificially inflate the gap. We address anchor sensitivity by running the same experiment with three alternative anchor sets and reporting Spearman rank correlation between the resulting score gaps (Appendix A). High correlation indicates the findings are not specific to our chosen anchors.

3.5. Residual Stream Patching

For each prompt pair we run the following procedure:

1. Run the model on the clean tokens, caching `hook_resid_pre` activations at every layer. We cache only residual stream pre-activations instead of all intermediate activations to reduce memory overhead.
2. Run the model on the corrupted tokens.
3. For each layer $l \in \{0, \ldots, L-1\}$, replace the corrupted residual stream at layer l with the cached clean activations and measure the resulting valence score.
4. Record the patch effect at each layer: the change in valence score from patching at layer l.

We record two summary statistics per prompt:

- `top_layer`: the layer index with the maximum patch effect, showing where the valence signal is most causally concentrated.
- `max_patch_effect`: the magnitude of the peak patch effect, reflecting how strongly a single layer drives the output.

The distribution of `top_layer` across all prompts is the primary evidence for the layer dissociation finding.

3.6. Valence Flip Test

The flip test directly tests whether the model tracks valence as a separable variable or simply recognizes topic.

For each model, we align good-news and negative-control pairs by index (they share the same corrupted baseline) and compute the score gap for each pair under both conditions. A flip occurs when:

$$\text{gap}_{\text{good_news}} > 0 \quad \text{and} \quad \text{gap}_{\text{negative_control}} < 0 \qquad (2)$$

for the same prompt index.

A model that recognizes only topic would produce correlated gaps across conditions, yielding a flip rate near 50%. A

model with a genuine valence representation would produce anti-correlated gaps, yielding a flip rate well above 50%. We report flip rate per model and use it as the primary evidence against the topic detection hypothesis.

3.7. Causal Steering

The steering experiment tests whether the layer identified by patching encodes valence as a manipulable linear direction or merely correlates with valence-charged inputs.

Direction extraction. For each model and condition, we identify the top causal layer as the median `top_layer` across all 100 prompts. We then extract a valence direction by computing the mean activation difference between clean and corrupted runs at that layer across 50 randomly sampled prompt pairs:

$$\vec{v} = \frac{1}{N}\sum_{i=1}^{N}\left(h^l_{\text{clean},i} - h^l_{\text{corr},i}\right), \qquad \hat{v} = \frac{\vec{v}}{\|\vec{v}\|} \tag{3}$$

where $h^l_{\cdot,i}$ denotes the last-token residual stream activation at layer l for the i-th prompt. Unit-normalizing ensures the direction is scale-invariant.

Steering neutral prompts. We construct 50 emotionally neutral prompts. These are factual context-free sentences with no emotional content (e.g. "I went to the library and borrowed a book today."). For each neutral prompt we apply the intervention:

$$h^l \leftarrow h^l + \alpha \cdot \hat{v} \tag{4}$$

at the last token position at layer l, sweeping $\alpha \in \{-20, -10, -5, 0, 5, 10, 20\}$. We measure the resulting valence score after the intervention and compute $\Delta = \text{score}_{\text{steered}} - \text{score}_{\text{base}}$.

4. Results

4.1. Models Track Valence Directionally

Before examining internal representations, we verify that all three models actually produce directionally correct responses to emotional content. We compute the valence score (Eq. 1) for each clean and corrupted prompt pair and measure sign accuracy: the fraction of prompts where the score gap has the correct sign. This is a necessary baseline because if a model cannot distinguish good news from bad news in its output distribution, there is nothing to explain mechanistically. Figure 1 shows the full score gap distributions across all 100 prompts per condition which confirms the separation is consistent and not driven by a few outliers.

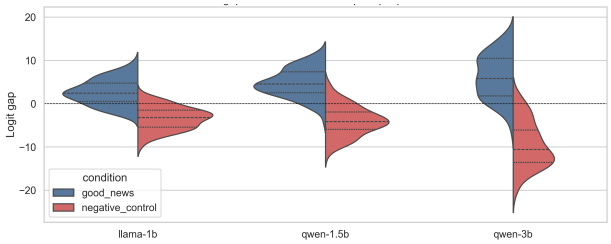

Figure 1. Score gap distributions across 100 prompts per condition for all three models. Each half-violin shows the distribution of logit gap scores: blue (left) for good-news prompts and red (right) for negative-control prompts. The two distributions are cleanly separated above and below zero in all three models.

Both model families respond correctly to valence across the majority of prompts (Table 1).

Model	Condition	Mean gap	Sign accuracy
Llama-1B	good_news	$+2.66$	79.0%
Llama-1B	negative_control	-3.39	88.0%
Qwen-1.5B	good_news	$+4.79$	92.0%
Qwen-1.5B	negative_control	-3.97	86.0%
Qwen-3B	good_news	$+6.27$	89.0%
Qwen-3B	negative_control	-9.45	95.0%

Table 1. Mean score gap (Eq. 1) and sign accuracy per model and condition. Positive gap = lean toward positive anchor tokens; negative gap = lean toward negative anchors.

Sign accuracy ranges from 79% to 95% across models and conditions. The Qwen family produces sharper score gaps than Llama-1B, with Qwen-3B showing the strongest negative-control signal (mean gap -9.45, sign accuracy 95%). Llama-1B's good-news sign accuracy of 79% is the weakest result in the table which suggests positive valence tracking is noisier in the smaller model.

4.2. The Flip Test Rules Out Topic Detection

High sign accuracy is a necessary but not sufficient check. A model that recognizes "PhD program" as emotionally loaded could produce a positive gap for both conditions if topic alone biases the output. The flip test directly targets this alternative by checking whether the score gap changes sign when valence changes while the topic and corrupted baseline are held fixed.

Valence flip rates are: Llama-1B 69%, Qwen-1.5B 80% and Qwen-3B 85%, all well above the 50% chance baseline. These rates are 19–35 percentage points above chance, making the margin substantial and unlikely to be attributable to noise. Sign accuracy tells us the model responds correctly to valence while the flip test tells us the response is driven specifically by valence and not by topic, syntax or any other confound shared across conditions. The consistency across Llama-1B (MHA) and Qwen models (GQA) further sug-

gests this is not an architecture-specific artifact. Together, these results establish that the model encodes valence as a separable internal variable, which is the precondition for the layer dissociation analysis that follows.

4.3. Negative and Positive Valence Localize at Different Depths

Having confirmed that models respond directionally to valence, we now ask where in the network this computation happens. For each prompt pair we run residual stream patching across all layers and record the layer with the peak patch effect (`top_layer`). If they shared a single representation, top layers would cluster at the same depth. A systematic difference indicates distinct computational pathways.

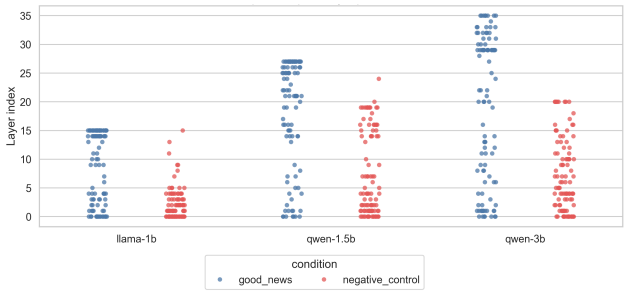

Figure 2. Top causal patch layer per prompt across all three models. Blue dots (good news) cluster at higher layers; red dots (negative control) cluster near the bottom. Each dot is one prompt; the dissociation is consistent and not driven by outliers. Good-news dots show wider vertical spread while negative-control dots pack tightly near the bottom, showing diffuse positive processing against sharply localized negative processing.

This is the central finding of the paper. Good-news signal peaks at 53–66% of model depth; negative-control signal peaks at 14–27%. Table 2 confirms this dissociation is highly significant across all three models.

Model	n_{layers}	U	p-value
Llama-1B	16	7808	1.84×10^{-12}
Qwen-1.5B	28	8071	2.58×10^{-14}
Qwen-3B	36	7501	4.62×10^{-10}

Table 2. Mann-Whitney U test on `top_layer` distributions: good_news vs. negative_control. One-sided alternative: good_news top layers exceed negative_control top layers. All three models are highly significant which confirms the early-late dissociation is not driven by chance or outliers.

Early layers in transformers are generally associated with low-level lexical processing (Elhage et al., 2021). The fact that negative valence peaks there suggests the model has a fast shallow detector for negative outcomes. Outcome words like "rejected" and "failed" are lexically distinctive and may not require deep contextual integration to trigger a response. Positive valence, by contrast, peaks at 53–66% of model depth. This suggests that producing a warm response requires the model to first assemble a richer contextual representation of what the good outcome actually means. This asymmetry is consistent across all three models regardless of whether they use MHA or GQA attention. Domain-level analysis (Appendix B) confirms the dissociation holds within academia, career and personal domains across all models ($p < 0.05$).

4.4. Patch Magnitude Tracks Response Strength Asymmetrically

Beyond identifying which layer peaks, we ask whether the size of the patch effect predicts how strongly the model responds. If valence is concentrated in one layer, a larger patch effect should produce a larger output shift. We measure Spearman rank correlation between `max_patch_effect` and score gap across all prompts per model and condition. A strong correlation indicates localized processing while a weak one indicates the signal is distributed across many layers.

Figure 3 shows this asymmetry visually. For negative control, patch magnitude strongly predicts response strength, with a single dominant layer driving the effect. The relationship collapses for good news indicative of positive valence being distributed across layers. Table 3 reports the Spearman rank correlation between patch magnitude and score gap per model and condition. A large negative ρ means that prompts where one layer dominates the patch effect also tend to have more extreme score gaps, pointing to tight causal localization. A ρ near zero means the two are unrelated, pointing to distributed processing.

Model	Condition	Spearman ρ	p-value
Llama-1B	good_news	$+0.144$	0.15
Llama-1B	negative_control	-0.646	$< 10^{-12}$
Qwen-1.5B	good_news	$+0.372$	1.37×10^{-4}
Qwen-1.5B	negative_control	-0.488	2.60×10^{-7}
Qwen-3B	good_news	$+0.112$	0.27
Qwen-3B	negative_control	-0.676	$< 10^{-13}$

Table 3. Spearman ρ between max patch effect and score gap. Strong negative correlation for negative control points to a single dominant layer driving the effect. Weak and non-significant correlation for good news reflects diffuse processing across layers.

For negative control, patch magnitude is a reliable predictor of response strength across all three models (ρ between -0.49 and -0.68, all $p < 10^{-6}$). This means that when the model reads bad news, a single dominant layer drives the output. Its patch magnitude predicts how negative the response will be. However, for good news, this relationship is absent in Llama-1B and Qwen-3B ($\rho = 0.14$ and 0.11) and only moderate in Qwen-1.5B ($\rho = 0.37$). Positive valence does not route through a single dominant layer. The

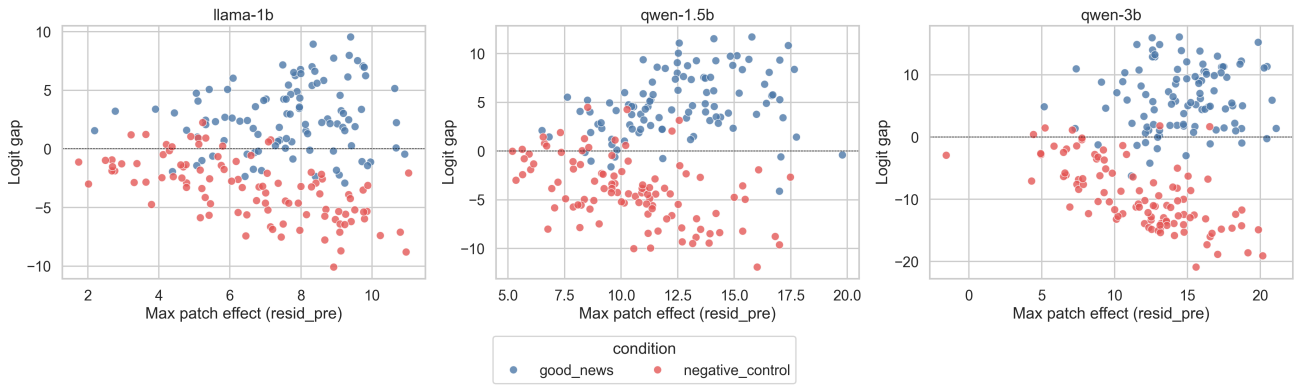

Figure 3. Max patch effect vs. score gap per prompt. For negative control (red), larger patch effects reliably predict more negative score gaps, with a single dominant layer driving the effect. For good news (blue) the relationship is weak, suggesting positive valence is more diffusely distributed across layers.

causal signal is spread across the network, which is why no single patch location strongly predicts response strength.

4.5. Steering Closes the Causal Loop

Patching shows which layer carries the valence signal but does not prove the residual stream at that layer encodes valence as a manipulable direction. A layer could be causally active because it routes information instead of storing it. To test this, we extract a mean-difference valence direction at the top causal layer for each condition (Eq. 3) and add it to the residual stream of 50 emotionally neutral prompts.

Steering at the patching-identified layers shifts neutral prompts toward positive valence when the good-news direction is added across all models (Table 4). The α-Δ relationship is shown in Figure 4.

Model	% shifted (+)	% shifted (-)	Mean Δ (+)
Llama-1B	100%	70%	+4.43
Qwen-1.5B	100%	82%	+3.07
Qwen-3B	92%	84%	+1.13

Table 4. Steering results on 50 neutral prompts using the good-news valence direction. (+) is the percentage of prompts that shift toward positive valence at $\alpha = +10$. (−) is the percentage that shift toward negative valence at $\alpha = -10$. Mean Δ is the average change in valence score at $\alpha = +10$.

At $\alpha = +10$, 100% of Llama-1B prompts and 100% of Qwen-1.5B prompts shift positive, with Qwen-3B at 92%. The valence score increases by $+4.43$ for Llama-1B, $+3.07$ for Qwen-1.5B and $+1.13$ for Qwen-3B at $\alpha = +10$ on prompts with no emotional content. At $\alpha = -10$, 70%, 82% and 84% of prompts shift toward negative valence for Llama-1B, Qwen-1.5B and Qwen-3B respectively. This shows that the good-news direction has genuine bidirectional control over emotional lean. The relationship between α and Δ is monotonically increasing for the good-news direction in all three models (Spearman $\rho > 0.89$, $p < 10^{-100}$). This confirms that the extracted direction is a genuine linear valence direction and not a noise vector.

5. Discussion

The early-late asymmetry. The dissociation is the most surprising result. It is loosely consistent with negativity bias in biological systems (Cacioppo et al., 1999; Baumeister et al., 2001) where threats are processed faster than rewards. A simpler explanation is that negative outcome words ("rejected," "failed," "denied") are more lexically distinctive and detectable from surface features alone, requiring less contextual integration than positive outcomes. Distinguishing these two explanations is an important direction for future work.

Implications. The finding that negative and positive valence are processed at different network depths has direct consequences for how we think about emotion in LLMs. It suggests these are not two ends of a single valence axis but distinct computational processes that can in principle be targeted independently. The early negative-valence signal in particular is a natural candidate for monitoring because it appears before the model has assembled a full response and is predictably steerable.

6. Limitations

Our metric is a proxy for valence sensitive to anchor choice. Stability across three anchor sets (Appendix A) with ρ between 0.47 and 0.87 indicates the findings are not strongly anchor-driven, though the lower end of this range leaves some residual uncertainty. The steering direction is extracted from the same prompt distribution used for patching

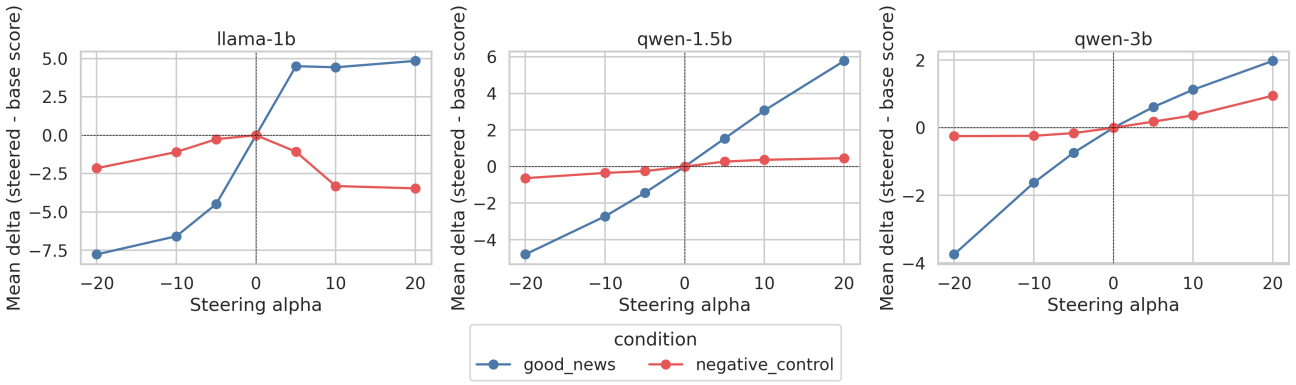

Figure 4. Mean change in valence score (Δ = steered minus base) as a function of steering strength α, averaged over 50 neutral prompts for each model. Blue line shows results for the good-news valence direction extracted at the good-news top layer. Red line shows results for the negative-control direction extracted at the negative-control top layer. The blue line rises monotonically confirming a clean steerable positive valence direction. The red line varies across models. Llama-1B shows a strongly inverted response while Qwen models show a weaker and flatter signal.

and may overfit to it. Future work should validate on out-of-distribution emotional content. We study instruct-tuned models and whether the same asymmetry exists in base models remains an open question. The Llama-1B vs. Qwen-1.5B comparison is approximate because the models differ in both architecture and parameter count.

7. Conclusion

We have shown that emotional valence in LLMs is not a monolithic internal state but an asymmetric computation. Negative valence localizes causally to early layers while positive valence peaks at mid-to-late layers. A controlled flip test rules out topic detection as an alternative explanation. A mean-difference valence direction extracted at those layers shifts the vast majority of emotionally neutral prompts toward positive valence, confirming the residual stream there encodes valence as a genuinely manipulable representation. Understanding why the two polarities require such different depths of computation may reveal something fundamental about how transformers encode semantic and emotional content.

Acknowledgements

The authors thank RunPod for GPU compute resources used to run all experiments in this work and Hugging Face for hosting the open-source models used throughout.

Impact Statement

This work studies the internal representations of emotional valence in open-source language models. We do not foresee direct negative societal impacts from this research. The findings may inform future work on monitoring internal model states, which could contribute to safer and more interpretable language model deployments. All models studied are publicly available and the experiments do not involve human subjects or sensitive data.

Disclosure of AI Usage

Large language models were used for language editing of the draft (Gemini) and code assistance (Claude Code) in this work. All research ideas, technical content, experimental design, results and conclusions are the original work of the authors.

References

Ainslie, J., Lee-Thorp, J., De Jong, M., Zemlyanskiy, Y., Lebrón, F., and Sanghai, S. Gqa: Training generalized multi-query transformer models from multi-head checkpoints. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 4895–4901, 2023.

Arditi, A., Obeso, O., Syed, A., Paleka, D., Panickssery, N., Gurnee, W., and Nanda, N. Refusal in language models is mediated by a single direction. *Advances in Neural Information Processing Systems*, 37:136037–136083, 2024.

Baumeister, R. F., Bratslavsky, E., Finkenauer, C., and Vohs, K. D. Bad is stronger than good. *Review of general psychology*, 5(4):323–370, 2001.

Bricken, T., Templeton, A., Batson, J., Chen, B., Jermyn, A., Conerly, T., Turner, N., Anil, C., Denison, C., Askell, A., Lasenby, R., Wu, Y., Kravec, S., Schiefer, N., Maxwell, T., Joseph, N., Hatfield-Dodds, Z., Tamkin, A., Nguyen, K., McLean, B., Burke, J. E., Hume, T., Carter, S., Henighan, T., and Olah, C. Towards monosemanticity:

Decomposing language models with dictionary learning. *Transformer Circuits Thread*, 2023. https://transformer-circuits.pub/2023/monosemantic-features/index.html.

Cacioppo, J. T., Gardner, W. L., and Berntson, G. G. The affect system has parallel and integrative processing components: Form follows function. *Journal of personality and Social Psychology*, 76(5):839, 1999.

Conmy, A., Mavor-Parker, A., Lynch, A., Heimersheim, S., and Garriga-Alonso, A. Towards automated circuit discovery for mechanistic interpretability. *Advances in Neural Information Processing Systems*, 36:16318–16352, 2023.

Cordonnier, J.-B., Loukas, A., and Jaggi, M. Multi-head attention: Collaborate instead of concatenate. *arXiv preprint arXiv:2006.16362*, 2020.

Elhage, N., Nanda, N., Olsson, C., Henighan, T., Joseph, N., Mann, B., Askell, A., Bai, Y., Chen, A., Conerly, T., et al. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 1(1):12, 2021.

Elhage, N., Hume, T., Olsson, C., Schiefer, N., Henighan, T., Kravec, S., Hatfield-Dodds, Z., Lasenby, R., Drain, D., Chen, C., et al. Toy models of superposition. *arXiv preprint arXiv:2209.10652*, 2022.

Hofmann, V., Kalluri, P. R., Jurafsky, D., and King, S. Dialect prejudice predicts ai decisions about people's character, employability, and criminality. *arXiv preprint arXiv:2403.00742*, 2024.

Lieberum, T., Rahtz, M., Kramár, J., Nanda, N., Irving, G., Shah, R., and Mikulik, V. Does circuit analysis interpretability scale? evidence from multiple choice capabilities in chinchilla. *arXiv preprint arXiv:2307.09458*, 2023.

Marks, S. and Tegmark, M. The geometry of truth: Emergent linear structure in large language model representations of true/false datasets. *arXiv preprint arXiv:2310.06824*, 2023.

Meng, K., Bau, D., Andonian, A., and Belinkov, Y. Locating and editing factual associations in gpt. *Advances in neural information processing systems*, 35:17359–17372, 2022.

Meta. Llama 3.2: Model cards and prompt formats. https://www.llama.com/docs/model-cards-and-prompt-formats/llama3_2/, 2024.

nostalgebraist. Interpreting GPT: the logit lens, 2020. URL https://www.lesswrong.com/posts/AcKRB8wDpdaN6v6ru/interpreting-gpt-the-logit-lens.

Olsson, C., Elhage, N., Nanda, N., Joseph, N., DasSarma, N., Henighan, T., Mann, B., Askell, A., Bai, Y., Chen, A., et al. In-context learning and induction heads. *arXiv preprint arXiv:2209.11895*, 2022.

Panickssery, N., Gabrieli, N., Schulz, J., Tong, M., Hubinger, E., and Turner, A. M. Steering llama 2 via contrastive activation addition. *arXiv preprint arXiv:2312.06681*, 2023.

Park, K., Choe, Y. J., and Veitch, V. The linear representation hypothesis and the geometry of large language models. *arXiv preprint arXiv:2311.03658*, 2023.

Sofroniew, N., Kauvar, I., Saunders, W., Chen, R., Henighan, T., Hydrie, S., Citro, C., Pearce, A., Tarng, J., Gurnee, W., et al. Emotion concepts and their function in a large language model. *arXiv preprint arXiv:2604.07729*, 2026.

Tak, A. N., Banayeeanzade, A., Bolourani, A., Kian, M., Jia, R., and Gratch, J. Mechanistic interpretability of emotion inference in large language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 13090–13120, 2025.

Templeton, A., Conerly, T., Marcus, J., Lindsey, J., Bricken, T., Chen, B., Pearce, A., Citro, C., Ameisen, E., Jones, A., Cunningham, H., Turner, N. L., McDougall, C., MacDiarmid, M., Freeman, C. D., Sumers, T. R., Rees, E., Batson, J., Jermyn, A., Carter, S., Olah, C., and Henighan, T. Scaling monosemanticity: Extracting interpretable features from claude 3 sonnet. *Transformer Circuits Thread*, 2024. URL https://transformer-circuits.pub/2024/scaling-monosemanticity/index.html.

Tigges, C., Hollinsworth, O. J., Geiger, A., and Nanda, N. Linear representations of sentiment in large language models. *arXiv preprint arXiv:2310.15154*, 2023.

Turner, A. M., Thiergart, L., Leech, G., Udell, D., Mini, U., and MacDiarmid, M. Activation addition: Steering language models without optimization. 2024.

Vig, J., Gehrmann, S., Belinkov, Y., Qian, S., Nevo, D., Sakenis, S., Huang, J., Singer, Y., and Shieber, S. Causal mediation analysis for interpreting neural nlp: The case of gender bias. *arXiv preprint arXiv:2004.12265*, 2020.

Wang, C., Zhang, Y., Yu, R., Zheng, Y., Gao, L., Song, Z., Xu, Z., Xia, G., Zhang, H., Zhao, D., et al. Do llms" feel"? emotion circuits discovery and control. *arXiv preprint arXiv:2510.11328*, 2025.

Wang, K., Variengien, A., Conmy, A., Shlegeris, B., and Steinhardt, J. Interpretability in the wild: a circuit for indirect object identification in gpt-2 small. *arXiv preprint arXiv:2211.00593*, 2022.

Yang, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Li, C., Liu, D., Huang, F., Wei, H., et al. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

Zhang, J. and Zhong, L. Decoding emotion in the deep: A systematic study of how llms represent, retain, and express emotion. *arXiv preprint arXiv:2510.04064*, 2025.

Zou, A., Phan, L., Chen, S., Campbell, J., Guo, P., Ren, R., Pan, A., Yin, X., Mazeika, M., Dombrowski, A.-K., et al. Representation engineering: A top-down approach to ai transparency. *arXiv preprint arXiv:2310.01405*, 2023.

A. Anchor Sensitivity

We verify the valence metric (Eq. 1) is stable across three anchor sets. The metric measures emotional lean by computing a logit gap between a fixed set of positive and negative response tokens. A concern with any such metric is that the findings could be specific to the particular tokens chosen. Different tokens might rank prompts differently, changing which prompts appear to have strong valence signal and which do not. To test this, we define two alternative anchor sets (Alt1, Alt2) that cover the same emotional territory using lexically distinct tokens and repeat the score gap computation for each. If the findings are genuine, score gaps computed with different anchor sets should rank prompts similarly, producing high Spearman ρ.

Table 5 reports both the mean score gap and Spearman ρ between each pair of anchor sets. Several patterns are worth noting. First, all three sets produce sign-consistent mean gaps throughout. Every model under good-news shows a positive mean gap and every model under negative-control shows a negative mean gap, regardless of which anchor set is used. The sign of the effect is therefore not anchor-dependent. Second, rank correlations between sets are moderate to strong (ρ between 0.47 and 0.87), with most values exceeding 0.6. The strongest correlations are between Alt1 and Alt2, suggesting these two sets are more similar to each other than either is to the default set. Third, the lowest correlations occur for Llama-1B good-news (default vs. Alt1: $\rho = 0.47$). This reflects the weaker overall good-news signal in Llama-1B, where noisier signals produce less stable rank orderings across anchor sets. Notably, Llama-1B's negative-control condition is more stable (ρ between 0.50 and 0.79), consistent with the stronger and more localized negative valence signal observed in the main results.

		Mean score gap			Spearman ρ		
Model	Condition	Default	Alt1	Alt2	D vs A1	D vs A2	A1 vs A2
Llama-1B	good_news	+2.74	+2.41	+3.01	0.647	0.744	0.867
Llama-1B	negative_control	−3.49	−5.15	−5.63	0.503	0.693	0.791
Qwen-1.5B	good_news	+4.82	+3.11	+4.98	0.472	0.690	0.768
Qwen-1.5B	negative_control	−3.96	−9.15	−7.68	0.631	0.585	0.840
Qwen-3B	good_news	+6.95	+6.25	+8.06	0.673	0.709	0.873
Qwen-3B	negative_control	−9.16	−20.19	−15.20	0.746	0.622	0.806

Table 5. Anchor sensitivity table showing mean score gap and Spearman ρ between all pairs of anchor sets per model and condition. Mean score gap columns confirm the sign of the effect is the same across all three anchor sets: positive for good news and negative for negative control throughout. Spearman ρ columns report rank correlations between sets (ρ between 0.47 and 0.87) which shows that prompt orderings are broadly stable across anchor choices.

The anchor sets used are:

- **Default:** congratulations, happy, glad, wonderful, amazing, thrilled, proud, fantastic, excellent | okay, noted, fine, ordinary, received, sorry
- **Alt1:** great, awesome, excited, delighted, superb, impressive, brilliant, pleased, overjoyed, lucky | unfortunately, sadly, regret, disappointing, difficult, tough
- **Alt2:** celebrate, congratulate, achievement, success, earned, deserve, outstanding, remarkable, triumph | failed, rejected, denied, missed, lost, unfortunate

B. Domain Analysis

The domain analysis tests whether the layer dissociation finding holds within each domain independently. If the finding were driven by topic-specific vocabulary, for example if academic words like "accepted" and "rejected" happened to produce early-layer responses due to their frequency in training data, we would expect the finding to disappear within career or personal prompts where the vocabulary is entirely different. Conversely, if the dissociation holds across all three domains, it provides strong evidence that the model is responding to valence as a general property and not to any single domain's lexical patterns.

Representative prompt pairs. Table 6 shows one clean/corrupted pair from each domain. The corrupted prompt is emotionally neutral and identical across good-news and negative-control conditions within each domain. This design ensures that any difference in patch effect between conditions within a domain is attributable to valence alone and not to phrasing

or topic differences. The prompts were written to be realistic and emotionally clear while avoiding artificially extreme language.

Domain	Clean prompt	Corrupted prompt
Academia	*Good:* "I just got accepted into my dream PhD program today." *Negative:* "I just got rejected from my dream PhD program today."	"I just received an email about my PhD program today."
Career	*Good:* "My contract was renewed with a significant pay increase." *Negative:* "My contract was not renewed and I was let go this week."	"My contract renewal came through this week."
Personal	*Good:* "My biopsy results came back completely benign." *Negative:* "My biopsy results came back positive and I need further treatment."	"I received my biopsy results from the clinic."

Table 6. Representative prompt pairs from each domain. The corrupted prompt is emotionally neutral and shared across both conditions. Only the clean prompt varies in valence.

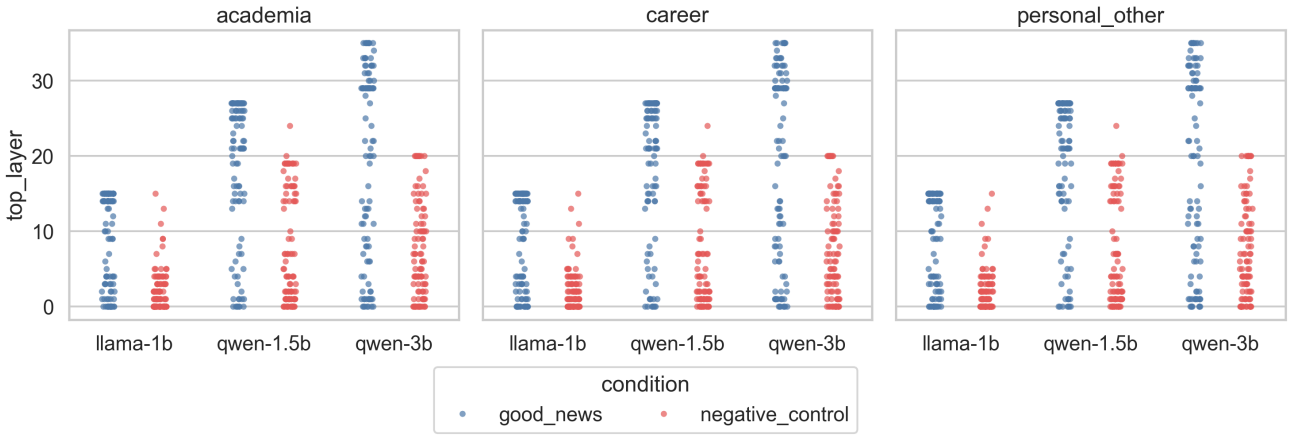

Figure 5. Top causal patch layer per prompt, broken down by domain. Blue dots (good news) cluster higher on the y-axis than red dots (negative control) within academia, career and personal domains independently. Each panel replicates the early-late dissociation seen in the main results within a single domain, ruling out the possibility that the finding is driven by vocabulary specific to any one topic area. The vertical spread within each condition reflects natural variation in which layer peaks for individual prompts.

Results. Table 7 reports Mann-Whitney U p-values for the layer dissociation within each domain. Academia and personal are highly significant across all three models ($p < 10^{-4}$), confirming the early-late dissociation is not confined to a single topic area. Career is significant for Qwen-1.5B ($p = 1.7 \times 10^{-3}$) and borderline for Llama-1B and Qwen-3B. The weaker career result is not surprising. Career outcome prompts tend to use more formal and hedged language than academia or personal prompts, which may reduce the clarity of the valence signal. The key result is that at least two of the three domains are independently significant in every model, ruling out the possibility that the main finding is driven by a single domain.

C. Patch Effect Visualizations

Residual stream patch heatmaps provide a per-prompt view of where the causal signal lives across layers and token positions. Each cell in the heatmap corresponds to one layer (y-axis) and one token position (x-axis). The value in each cell is the patch effect at that layer and position: how much the valence score changes when the clean activation at that layer and position is patched into the corrupted run. Brighter (yellow/green) cells indicate larger positive patch effects. Darker (blue/purple) cells indicate smaller or negative effects.

The key comparison is between the good-news and negative-control heatmaps for the same model. The layer dissociation

Domain	Llama-1B	Qwen-1.5B	Qwen-3B
academia	7.9×10^{-6}	6.3×10^{-5}	9.7×10^{-5}
career	3.2×10^{-2}	1.7×10^{-3}	6.9×10^{-2}
personal	1.9×10^{-7}	3.7×10^{-8}	6.5×10^{-8}

Table 7. Mann-Whitney U p-values testing whether good-news top layers are significantly higher than negative-control top layers within each domain. Academia and personal are highly significant across all three models ($p < 10^{-4}$), confirming the early-late dissociation holds within each domain independently. Career is significant for Qwen-1.5B and borderline for Llama-1B and Qwen-3B, likely due to more hedged language in career prompts reducing the clarity of the valence signal.

predicts that good-news signal should concentrate at mid-to-late layers while negative-control signal should concentrate near the bottom of the network. The token position axis also provides interpretable structure, with signal tending to concentrate around the outcome phrase instead of around neutral context tokens. We show one personal-domain prompt pair per model below.

Qwen-1.5B Inputs (Personal)

Good news: "I passed my professional certification exam on the first attempt."

Negative control: "I failed my language proficiency certification after months of preparation."

Corrupted: "I completed my professional certification exam today."

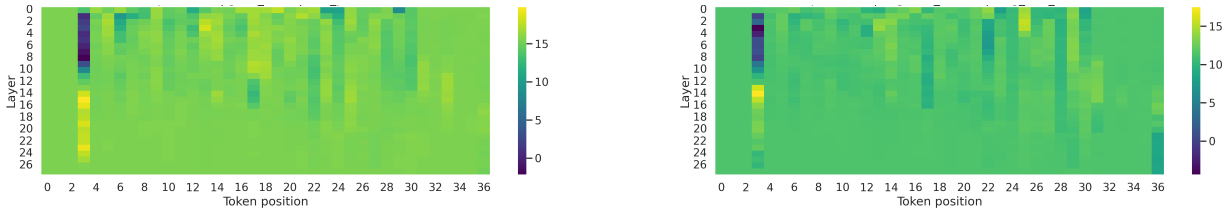

Figure 6. Qwen-1.5B residual stream patch heatmaps. Good news (left) shows signal spread across mid-to-late layers. Negative control (right) shows activity concentrated in the top rows (early layers), matching the layer dissociation pattern observed in Qwen-3B.

Qwen-3B Inputs (Personal)

Good news: "I passed my professional certification exam on the first attempt."

Negative control: "I was rejected from my first-choice university with no feedback."

Corrupted: "I completed my professional certification exam today."

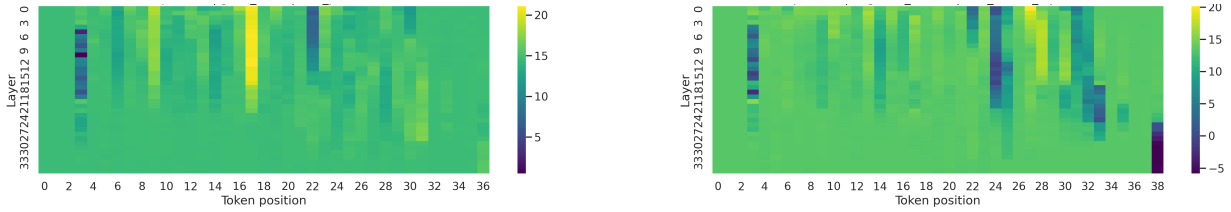

Figure 7. Qwen-3B residual stream patch heatmaps. Good news (left) shows bright cells concentrated around the outcome phrase tokens at mid-depth layers. Negative control (right) shows stronger activity in the top rows (early layers), directly visualizing the layer dissociation.

Llama-1B Inputs (Personal)

Good news: "I completed the hardest hiking trail in my region without stopping."

Negative control: "I gained back all the weight I had lost over six months."

Corrupted: "I completed a long hiking trail in my region today."

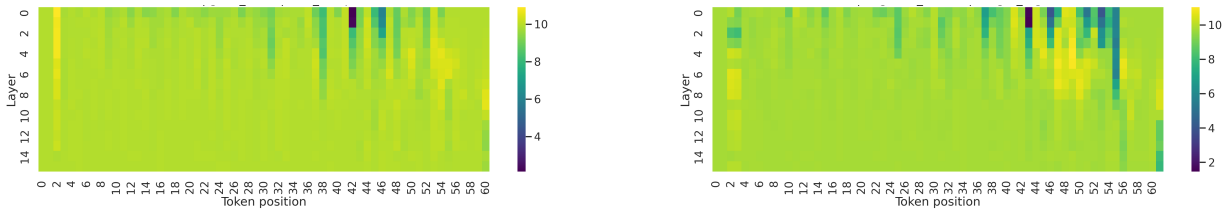

Figure 8. Llama-1B residual stream patch heatmaps. The negative-control heatmap (right) shows stronger activity in early layers compared to the good-news heatmap (left), where signal is more broadly distributed across the full depth of the network. The same early-late asymmetry holds as in Qwen-3B despite the architectural difference.