A Mechanistic Study of Mental Health Disclosure in Language Models

Anonymous ACL submission

Abstract

Large language models (LLMs) are increasingly deployed in mental health contexts, yet little is known about how they internally represent clinical states such as anxiety, depression, and suicidal ideation. We introduce a clinically grounded interpretability pipeline that measures each condition with a validated self-report instrument and combines this with mechanistic interventions on internal activations across seven open-weight LLMs. We find that all three conditions are encoded in a contiguous band of mid-to-late residual-stream layers, and that injecting a learned direction at these layers produces graded, predictable shifts in psychometric scores. These findings offer a direct window into how LLMs represent mental health concepts, with implications for the safer design of clinical LLM applications.

1 Introduction

The deployment of large language models (LLMs) as conversational agents is rapidly expanding across mental health, spanning emotional support, clinical screening, and therapeutic interactions (Jin et al., 2025; Lawrence et al., 2024; Heinz et al., 2025). Recent work has documented that these models can exhibit bias and sycophancy and reinforce distorted beliefs when prompted with emotionally loaded content (Sharma et al., 2024; Moore et al., 2025; Huang et al., 2025). Yet the internal mechanisms driving such behaviors remain opaque (Bereska and Gavves, 2024). It is imperative to understand how LLMs process and respond to such disclosures.

Mechanistic interpretability and representation engineering address this gap by identifying which internal components of a model cause specific behaviors (Olah et al., 2020; Vig et al., 2020). These methods combine observational tools (Hewitt and Manning, 2019; Cunningham et al., 2023) with interventional tools like activation patching and steering (Heimersheim and Nanda, 2024; Zou et al., 2023). Traditionally, these methods have been applied to linguistic, factual, reasoning, and bias circuits (Meng et al., 2023; Nanda et al., 2023; Olsson et al., 2022; Ahsan et al., 2025).Recent work extends this lens to emotion, finding that affective representations sit in mid-layer attention heads and structured subspaces, generalizing across model families (Tak et al., 2025; Zhang and Zhong, 2025; Wang et al., 2025; Sofroniew et al., 2026). However, while existing work focuses on emotions like joy and anger,whether the same techniques recover clinically meaningful mental health constructs remains unexplored.

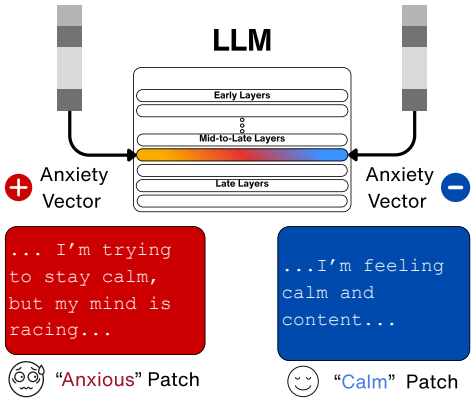

Figure 1: The figure illustrates activation steering via vector addition and subtraction inside a language model.

To address this gap, we investigate whether LLMs form distinct internal representations of anxiety, depression, and suicidal ideation (hereafter *constructs*, following standard psychometric usage), and evaluate whether targeted interventions on these representations can modify model behavior. To this end, our contributions are twofold: (1) **Clinically grounded interpretability pipeline:** We introduce a unified pipeline that pairs three validated Likert-scale self-report instruments with mechanistic interpretability techniques that include activation patching and steering (Figure 1). We apply this pipeline across seven LLMs to characterize

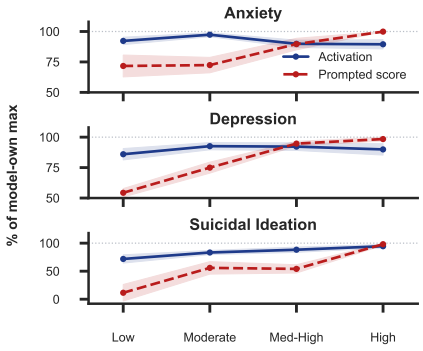

Figure 2: Internal activations saturate faster than prompted behavioral scores. The solid blue line is the mean residual-stream ℓ_2 activation distance from baseline; the dashed red line is the mean prompted psychometric score (STICSA, PHQ-9, or SIDAS) (without steering). For each model, both quantities are first normalized to that model's own maximum across the five intensity levels (100 on the y-axis corresponds to each model's peak value, not a shared absolute scale)

where mental health related information is encoded in model activations. (2) **Model intervention analysis:** We examine whether targeted activation-level interventions can causally modulate model behavior on validated clinical symptom measures. Our findings have implications for both the science of LLM psychological representations and their safe deployment in mental health contexts.

2 Methodology

2.1 Overview

For each of the three constructs, we generate 500 first-person stories at five affective intensity levels and extract activations at three sublayer locations. We then apply two interventions in sequence: activation patching identifies which sites are responsible for construct expression (Section 2.3), and activation steering then injects a vector at those sites to test whether we can modulate the constructs. We score each construct using next-token logits over the questionnaire response options (Section 2.4).

2.2 Psychometric Tests

Each construct is measured via a validated self-report instrument consisting of Likert-scale *items* (e.g., *"My thoughts are racing"*). Specifically, we use the STICSA (Ree et al., 2008) for anxiety, PHQ-9 (Kroenke et al., 2001) for depression, and SIDAS (Van Spijker et al., 2014) for suicidal ideation; items and scoring details are in Appendix I.

2.3 Localizing Constructs via Activation Patching

Preprompt Generation. We prompt the model to generate its own first-person story at a positively-valenced *baseline* (e.g., calm and mindful for anxiety) and four induced-affects (*low*, *moderate*, *medium-high*, *high*), following Coda-Forno et al. (2023) and the role-play framework of Shanahan et al. (2023). The full prompts are in Appendix I. We generate 100 stories per intensity.

Logit-Based Scoring. For each story, we append a psychometric item and score the model's response probabilistically by restricting next-token logits to the K Likert response option tokens and computing the expected ordinal score (Appendix F). To control for positional bias, each item is also administered with the option order reversed (Zheng et al., 2024).

Residual Stream Extraction. For each story we record activations at three points in every transformer block: the MHSA (Multi-Head Self Attention) output, the FFN (Feed Forward Network) output, and the residual stream. Activations are captured at the final token position (Belrose et al., 2023) and averaged across preprompt variations.

Activation Patching. This framework (Vig et al., 2020; Zhang and Nanda, 2024) identifies which internal sites drive a behavioral difference between two inputs. Treating the high-intensity condition as the *source* and the baseline as the *target*, we run the target forward pass while replacing a single activation (at layer ℓ, component c, and token t) with its cached source counterpart. Following Meng et al. (2022), we measure the resulting shift as the causal effect (CE), substituting expected questionnaire scores for token probabilities:

$$\mathrm{CE}_{\ell,c} = \frac{s_{\text{patched}}(\ell, c) - s_{\text{baseline}}}{s_{\text{source}} - s_{\text{baseline}}}, \qquad (1)$$

where s_{baseline}, s_{source}, and $s_{\text{patched}}(\ell, c)$ are the expected scores under each condition. $\mathrm{CE} = 0$ means the patch has no effect; $\mathrm{CE} = 1$ means it fully reproduces the source-induced behavior.

2.4 Manipulating Constructs via Activation Steering

After localizing the construct, we extract its direction and use it to causally manipulate the behavior.

Vector Direction. We define the construct direction as a unit vector $\mathbf{v}_{\ell,c,t} \in \mathbb{R}^d$, where positive

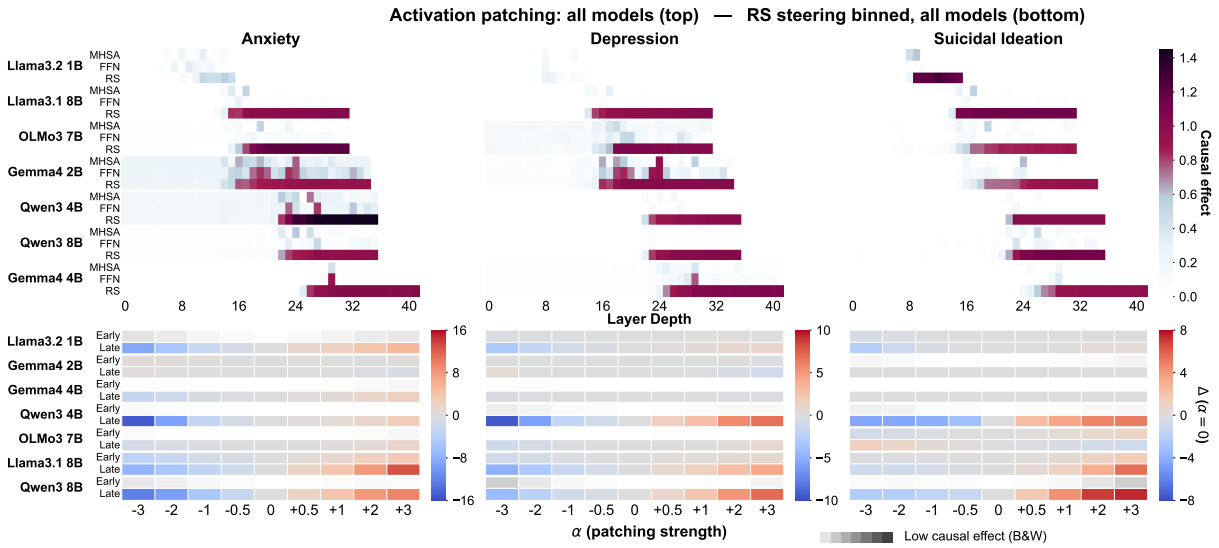

Figure 3: Activation patching (top) and residual-stream steering (bottom) across constructs. **Top:** causal effect per layer (x) and component (MHSA/FFN/RS, y);. **Bottom:** Δ score from $\alpha = 0$ under steering strength α, binned to early/late layers; grey = causal effect below 40% of model max. (top) localizes where each construct is represented (bottom) tests whether steering at those layers shifts the model's behavior.

displacements increase the construct intensity and negative displacements decrease it. Following Zou et al. (2023); Rimsky et al. (2024), we compute $\mathbf{v}$ as the normalized mean activation difference between high-intensity and baseline conditions:

$$\mathbf{v}_{\ell,c,t} = \frac{\bar{h}^{\text{high}}_{\ell,c,t} - \bar{h}^{\text{baseline}}_{\ell,c,t}}{\left\| \bar{h}^{\text{high}}_{\ell,c,t} - \bar{h}^{\text{baseline}}_{\ell,c,t} \right\|}, \tag{2}$$

where $\bar{h}$ denotes the mean activation vector across psychometric items.

Activation Steering. We administer the psychometric tests while registering forward hooks that modify activations during the forward pass. The activation is updated as

$$h'_{\ell,c,t} = h_{\ell,c,t} + \alpha \cdot \mathbf{v}^{\text{norm}}_{\ell,c,t}, \tag{3}$$

where $\alpha \in \mathbb{R}$ is a scalar gain: positive α elevates the construct, negative α suppresses it. To pool evidence across nearby layers, we apply steering simultaneously to a three-layer window centered on ℓ, namely $\{\ell - 1, \ell, \ell + 1\}$. To keep α comparable across layers, we rescale each direction so that $\alpha = 1.0$ corresponds to 5% of the mean baseline activation norm at that location (Appendix G). The tests are also cross-applied across constructs.

3 Experimental Setup

Our implementation extends the codebase of Tak et al. (2025), with anxiety-induction prompts adapted from Coda-Forno et al. (2023).

Models. We evaluate seven LLMs: Gemma 4 E2B-it and E4B-it (Google DeepMind, 2026), Llama 3.2 1B and Llama 3.1 8B Instruct (Grattafiori et al., 2024), OLMo 3 7B Instruct (Olmo et al., 2025), and Qwen 3 4B and Qwen 3 8B Instruct (Yang et al., 2025). Implementation details and selection rationale are in Appendix A.

Metrics. We evaluate interventions along two axes: *construct-specific behavior* and *output coherence*. Behavior is measured via the logit-based psychometric score S (Equation 4); for activation patching we additionally report the causal effect $\text{CE}_{\ell,c}$ (Equation 1). Coherence is measured via token-level perplexity and entropy on free-form responses after steering.

Ablations As a null control, we extract directions from semantically-unrelated contrast pairs and run them through the same alpha sweep (Appendix C).

4 Results

Even when prompted scores appear mild, the model's internal representations are already fully displaced. Figure 2 compares activation distance and psychometric scores under increasing prompt intensity. Both are normalized to each model's peak before averaging across the seven evaluated models (shaded bands: ± 1 SE). Across all three constructs, activations reach 90–100% of their peak at the lowest intensity while prompted scores remain at 50–60%, with the gap largest for

Model	α	Anxiety				Depression				Suicidal Ideation			
		Causal / Steer Site	$\Delta_{+\alpha}$	$\Delta_{-\alpha}$	Spec.	Causal / Steer Site	$\Delta_{+\alpha}$	$\Delta_{-\alpha}$	Spec.	Causal / Steer Site	$\Delta_{+\alpha}$	$\Delta_{-\alpha}$	Spec.
Llama3.2 1B	±3	RS L14 / RS L11	$+7.9$	-13.2	-4.7	RS L8 / RS L14	$+3.3$	-4.8	$+3.3$	RS L12 / RS L12	$+0.5$	-2.0	-1.3
Gemma4 2B	±2	RS L28 / FFN L1	$+4.1$	$+0.2$	-3.4	RS L34 / FFN L15	$+0.2$	-0.0	-2.4	RS L34 / RS L16	$+1.2$	$+1.4$	-6.2
Gemma4 4B	±3	RS L32 / RS L25	$+7.0$	-7.1	-3.9	RS L41 / RS L26	$+3.5$	-1.9	$+2.7$	RS L41 / RS L24	$+0.5$	-0.1	-2.2
Qwen3 4B	±3	RS L32 / RS L26	$+8.5$	-22.9	-45.2	RS L35 / RS L28	$+9.3$	-11.0	$+26.0$	RS L31 / RS L22	$+5.4$	-4.6	$+38.6$
OLMo3 7B	±3	RS L20 / MRSA L16	$+3.7$	-2.8	-4.3	RS L21 / RS L21	$+1.3$	-2.1	-8.4	RS L30 / RS L15	$+2.6$	-1.7	$+9.6$
Llama3.1 8B	±3	RS L29 / RS L15	$+15.8$	-16.5	-20.9	RS L31 / RS L23	$+4.7$	-4.4	-1.5	RS L31 / RS L15	$+8.1$	-1.2	$+24.7$
Qwen3 8B	$\pm3^*$	RS L27 / RS L24	$+10.5$	-14.5	$+1.9$	RS L34 / RS L26	$+10.2$	-3.7	-11.5	RS L30 / RS L24	$+7.6$	-2.2	$+24.8$

Table 1: Per-model causal localization and steering results. *Causal site* is the layer/component with the largest patching effect; *steer site* is the layer/component where steering has the strongest effect. $\Delta_{+\alpha}/\Delta_{-\alpha}$ are score changes at the listed gain (RS = residual stream). *Spec. margin* (%) is diagonal minus maximum off-target steering effect; positive values indicate construct-specific control.

	PHQ-9	STICSA	SIDAS
v_{dep}	2.1 ± 0.96	2.2 ± 0.75	0.3 ± 0.31
v_{anx}	-0.3 ± 0.56	1.0 ± 1.23	-0.4 ± 0.77
v_{si}	0.9 ± 0.52	1.7 ± 1.63	1.8 ± 0.74

Table 2: Cross-construct steering after projecting out $v_{\text{shared}} = \text{mean}(v_{\text{dep}}, v_{\text{anx}}, v_{\text{si}})$. Off-diagonals are reduced toward zero.

suicidal ideation (≈ 60 percentage points at low intensity). A model scoring in the mild range may therefore already carry a fully displaced internal representation, making standalone prompted scores an unreliable proxy for internal state.

The construct signal is encoded in the mid-to-late residual stream. Successful patching is sharply localized to a contiguous band of mid-to-late layers (Figure 3), with FFN patches peaking slightly later than MHSA patches at the same site. We interpret this as evidence that the construct signal is incrementally written into the residual stream by a contiguous block of MHSA-then-FFN computations. The causal band sits later for suicidal ideation than for anxiety, suggesting that more complex constructs require additional layers of compositional processing before being expressed.

Injecting the learned steering vector increases or decreases expressed anxiety, depression, and suicidal ideation. Steering at the identified mid-to-late layers produces a graded, monotonic shift in psychometric scores as a function of α (Figure 3): positive α raises the score and negative α lowers it, with both shifts scaling with $|\alpha|$. The effect is sharply localized, mirroring the patching band.

Per-model interventions reveal heterogeneous specificity. Table 1 shows that the largest positive shifts occur in larger dense models (Llama 3.1 8B anxiety, $\Delta_{+\alpha} = +15.8$; Qwen 3 8B depression $+10.2$ and suicidal ideation $+7.6$). While steering enables strong bidirectional score control, construct-specific control varies. Suicidal ideation separates most cleanly from the shared distress axis with the highest specificity, likely due to its distinctive cognitive and behavioral content (e.g., hopelessness, burdensomeness). Conversely, anxiety remains the most entangled.

Steered completions remain fluent. Median perplexity of free-form generations is flat across the steering sweep (no model exceeds a $1.07\times$ worst-case increase at $\alpha = \pm3$). Fluency degradation does not explain the observed score shifts (Appendix D).

Construct directions share a general-distress component. Each direction produces the largest effect on its target instrument but also smaller off-target shifts on the others (Table 2). These cross-talk effects trace to a shared direction with substantial layer-wise cosine similarity across constructs; projecting it out makes the steering matrix more diagonal (Appendix B).

5 Conclusion

We show that LLMs encode anxiety, depression, and suicidal ideation as causally manipulable representations in a contiguous band of mid-to-late residual-stream layers. Steering these representations produces graded, bidirectional shifts in psychometric scores. Internal activations frequently saturate before behavioral scores reflect the induced state. The three directions share a general-distress component while retaining separable construct-specific ones. Two implications follow. Prompted behavioral scores alone may miss internal state changes, a consequential blind spot for models deployed in screening or support contexts. And because the same directions can suppress as well as amplify a construct, steering offers a potential lever for mitigating affect-related failure modes.

Limitations

First of all, generalization beyond the seven models, three constructs, and the English language Psychometric instruments evaluated here should not be assumed. The three constructs themselves are also not fully separable: projecting out the shared component across v_{dep}, v_{anx}, and v_{si} makes the cross-construct steering matrix substantially more diagonal (Appendix B), indicating that construct-specific control requires explicit orthogonalization rather than emerging from the raw mean-difference direction. Secondly, our psychometric scoring measures the probability mass that the model places on each ordinal Likert option. This is a reproducible, continuous signal, but it is not equivalent to a clinical measurement on human respondents, and should be read as a probabilistic measure of model behavior on a structured instrument. Designing tighter contrasts that decouple constructs from each other, and validating logit-based ordinal scoring against human-administered protocols, are important directions for future work.

Ethics Statement

This work involves no human participants, private clinical records, or real-user evaluations; all psychometric responses are generated by large language models, and IRB review is therefore not required. Although the model is prompted to generate first-person descriptions of negative affective states, all such content is produced by the model itself, and no human-authored sensitive content is used or distributed. Psychometric scores are used as controlled measurements of model outputs, not as diagnoses or as evidence that models experience mental states.

Moreover, our work uses Mechanistic Interpretiblity and activation steering to localize and manipulate internal representations of clinical constructs, specifically anxiety, depression, and suicidal ideation. Our primary motivation is to advance safety-critical transparency and auditability in LLMs, though we acknowledge the dual-use risks inherent to activation-level modifications. Suppressing these constructs could mitigate harmful model behaviors, such as the reinforcement of distorted beliefs or sycophancy. Conversely, the same pipeline could be exploited to intentionally amplify or induce states of high anxiety, depression, or suicidal ideation in conversational agents.

References

Hiba Ahsan, Arnab Sen Sharma, Silvio Amir, David Bau, and Byron C. Wallace. 2025. Elucidating mechanisms of demographic bias in llms for healthcare. *Preprint*, arXiv:2502.13319.

Nora Belrose, Igor Ostrovsky, Lev McKinney, Zach Furman, Logan Smith, Danny Halawi, Stella Biderman, and Jacob Steinhardt. 2023. Eliciting latent predictions from transformers with the tuned lens. *arXiv preprint arXiv:2303.08112*.

Leonard Bereska and Efstratios Gavves. 2024. Mechanistic interpretability for ai safety–a review. *arXiv preprint arXiv:2404.14082*.

Julian Coda-Forno, Kristin Witte, Akshay K Jagadish, Marcel Binz, Zeynep Akata, and Eric Schulz. 2023. Inducing anxiety in large language models can induce bias. *arXiv preprint arXiv:2304.11111*.

Hoagy Cunningham, Aidan Ewart, Logan Riggs, Robert Huben, and Lee Sharkey. 2023. Sparse autoencoders find highly interpretable features in language models. *Preprint*, arXiv:2309.08600.

Google DeepMind. 2026. Gemma 4 model card. https://ai.google.dev/gemma/docs/core/model_card_4. Released April 2, 2026.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Phillip Guo, Aaquib Syed, Abhay Sheshadri, Aidan Ewart, and Gintare Karolina Dziugaite. 2024. Mechanistic unlearning: Robust knowledge unlearning and editing via mechanistic localization. *Preprint*, arXiv:2410.12949.

Stefan Heimersheim and Neel Nanda. 2024. How to use and interpret activation patching. *Preprint*, arXiv:2404.15255.

Michael V Heinz, Daniel M Mackin, Brianna M Trudeau, Sukanya Bhattacharya, Yinzhou Wang, Haley A Banta, Abi D Jewett, Abigail J Salzhauer, Tess Z Griffin, and Nicholas C Jacobson. 2025. Randomized trial of a generative ai chatbot for mental health treatment. *Nejm Ai*, 2(4):AIoa2400802.

John Hewitt and Christopher D. Manning. 2019. A structural probe for finding syntax in word representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4129–4138, Minneapolis, Minnesota. Association for Computational Linguistics.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen,

Weihua Peng, Xiaocheng Feng, Bing Qin, and 1 others. 2025. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2):1–55.

Y Jin, J Liu, P Li, B Wang, Y Yan, H Zhang, C Ni, J Wang, Y Li, Y Bu, and Y Wang. 2025. The applications of large language models in mental health: Scoping review. *Journal of Medical Internet Research*, 27:e69284.

Thomas E. Joiner. 2005. *Why People Die by Suicide*. Harvard University Press, Cambridge, MA.

Kurt Kroenke, Robert L Spitzer, and Janet BW Williams. 2001. The phq-9: validity of a brief depression severity measure. *Journal of general internal medicine*, 16(9):606–613.

Hannah R Lawrence, Renee A Schneider, Susan B Rubin, Maja J Matarić, Daniel J McDuff, and Megan Jones Bell. 2024. The opportunities and risks of large language models in mental health. *JMIR Mental Health*, 11(1):e59479.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022. Locating and editing factual associations in gpt. *Advances in neural information processing systems*, 35:17359–17372.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2023. Locating and editing factual associations in gpt. *Preprint*, arXiv:2202.05262.

Jared Moore, Declan Grabb, William Agnew, Kevin Klyman, Stevie Chancellor, Desmond C Ong, and Nick Haber. 2025. Expressing stigma and inappropriate responses prevents llms from safely replacing mental health providers. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*, pages 599–627.

Neel Nanda, Lawrence Chan, Tom Lieberum, Jess Smith, and Jacob Steinhardt. 2023. Progress measures for grokking via mechanistic interpretability. *Preprint*, arXiv:2301.05217.

Chris Olah, Nick Cammarata, Ludwig Schubert, Gabriel Goh, Michael Petrov, and Shan Carter. 2020. Zoom in: An introduction to circuits. *Distill*.

Team Olmo, Allyson Ettinger, Amanda Bertsch, Bailey Kuehl, David Graham, David Heineman, Dirk Groeneveld, Faeze Brahman, Finbarr Timbers, Hamish Ivison, and 1 others. 2025. Olmo 3. *arXiv preprint arXiv:2512.13961*.

Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Scott Johnston, Andy Jones, Jackson Kernion, Liane Lovitt, and 7 others. 2022. In-context learning and induction heads. *Preprint*, arXiv:2209.11895.

Melissa J Ree, Davina French, Colin MacLeod, and Vance Locke. 2008. Distinguishing cognitive and somatic dimensions of state and trait anxiety: Development and validation of the state-trait inventory for cognitive and somatic anxiety (sticsa). *Behavioural and Cognitive Psychotherapy*, 36(3):313–332.

Nina Rimsky, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Turner. 2024. Steering llama 2 via contrastive activation addition. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15504–15522.

Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. 2023. Whose opinions do language models reflect? In *International conference on machine learning*, pages 29971–30004. PMLR.

Nino Scherrer, Claudia Shi, Amir Feder, and David Blei. 2023. Evaluating the moral beliefs encoded in llms. *Advances in Neural Information Processing Systems*, 36:51778–51809.

Murray Shanahan, Kyle McDonell, and Laria Reynolds. 2023. Role play with large language models. *Nature*, 623(7987):493–498.

Mrinank Sharma, Meg Tong, Tomek Korbak, David Duvenaud, Amanda Askell, Sam Bowman, Esin Durmus, Zac Hatfield-Dodds, Scott Johnston, Shauna Kravec, and 1 others. 2024. Towards understanding sycophancy in language models. In *International Conference on Learning Representations*, volume 2024, pages 110–144.

Nicholas Sofroniew, Isaac Kauvar, William Saunders, Runjin Chen, Tom Henighan, Sasha Hydrie, Craig Citro, Adam Pearce, Julius Tarng, Wes Gurnee, Joshua Batson, Sam Zimmerman, Kelley Rivoire, Kyle Fish, Chris Olah, and Jack Lindsey. 2026. Emotion concepts and their function in a large language model. *Transformer Circuits Thread*.

Ala N. Tak, Amin Banayeeanzade, Anahita Bolourani, Mina Kian, Robin Jia, and Jonathan Gratch. 2025. Mechanistic interpretability of emotion inference in large language models. *Preprint*, arXiv:2502.05489.

Bregje AJ Van Spijker, Philip J Batterham, Alison L Calear, Louise Farrer, Helen Christensen, Julia Reynolds, and Ad JFM Kerkhof. 2014. The suicidal ideation attributes scale (sidas): Community-based validation study of a new scale for the measurement of suicidal ideation. *Suicide and Life-Threatening Behavior*, 44(4):408–419.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Simas Sakenis, Jason Huang, Yaron Singer, and Stuart Shieber. 2020. Causal mediation analysis for interpreting neural nlp: The case of gender bias. *Preprint*, arXiv:2004.12265.

Chenxi Wang, Yixuan Zhang, Ruiji Yu, Yufei Zheng, Lang Gao, Zirui Song, Zixiang Xu, Gus Xia,

Huishuai Zhang, Dongyan Zhao, and Xiuying Chen. 2025. Do llms "feel"? emotion circuits discovery and control. *Preprint*, arXiv:2510.11328.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Fred Zhang and Neel Nanda. 2024. Towards best practices of activation patching in language models: Metrics and methods. In *International Conference on Learning Representations*, volume 2024, pages 1651–1678.

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

Lining Zhang, Mengchen Wang, Liben Chen, and Wenxin Zhang. 2022. Probing GPT-3's linguistic knowledge on semantic tasks. In *Proceedings of the Fifth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 297–304, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Chujie Zheng, Hao Zhou, Fandong Meng, Jie Zhou, and Minlie Huang. 2024. Large language models are not robust multiple choice selectors. In *International Conference on Learning Representations*, volume 2024, pages 19426–19454.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, and 1 others. 2023. Representation engineering: A top-down approach to ai transparency. *arXiv preprint arXiv:2310.01405*.

A Implementation Details

Model selection. Our seven models span 1B–14B parameters, mix dense and mixture-of-experts (MoE) architectures, and draw from four distinct families (Meta, Google, AI2, Alibaba), enabling us to assess generalization across scale, architecture, and training lineage. For MoE models, interventions act on the post-routing residual stream to ensure consistent semantics across the full model set.

Preprompt generation. For each (construct, intensity) pair, the model generates $V = 100$ stories at temperature 1.0, top-p 0.95, and top-k 64, yielding 1,500 stories per model across the three constructs.

Activation extraction and intervention. Steering uses a sliding window of three consecutive layers $\{\ell - 1, \ell, \ell + 1\}$ (ablations at window sizes 1 and 5), with gain $\alpha \in \{-3, -2, -1, -0.5, 0, 0.5, 1, 2, 3\}$ rescaled per layer so that $\alpha = 1.0$ corresponds to 5% of the mean baseline activation norm (Equation 5).

Compute. 4× NVIDIA RTX A6000 GPUs (48 GB each), with the largest checkpoints sharded across cards. Total compute is approximately 80 GPU-hours, dominated by activation extraction across the 7 models × 3 constructs × 5 intensities × 100 stories generation grid and the steering sweep over 9 values of α applied to 3-layer windows at each candidate site.

Use of AI Assistants

We used LLM-based assistants (e.g., ChatGPT, Claude) for (i) writing assistance, including grammar correction, phrasing, and editing of draft text, and (ii) coding assistance, including boilerplate generation, refactoring, and debugging of the experiment and analysis code. All research ideas, experimental design, scientific claims, and final text were authored and verified by the human authors.

B Construct Specificity and Direction Geometry

Construct directions are partly specific and partly shared. Each construct direction produces its largest effect on its target instrument but also non-trivial off-target shifts (Table 3). At every residual-stream layer, the three directions share substantial cosine similarity ($\cos\mathrm{sim} \approx 0.3$–$0.5$; Figure 4). Projecting out $v_{\text{shared}} = \mathrm{mean}(v_{\text{dep}}, v_{\text{anx}}, v_{\text{si}})$ and re-steering makes the matrix substantially more diagonal (Table 2), indicating that the original cross-talk was driven by a single shared additive component rather than by deeper subspace overlap.

High-causal directions organize by construct. Projecting linear probe weight vectors at high-causal sites ($\mathrm{CE}_{\ell,c} \geq 0.7$) into the top two principal components reveals distinct construct clusters across all seven models. In the larger models (Llama 3.1 8B, OLMo 3 7B, Qwen 3 4B, Qwen 3 8B), the three clusters trace C-shaped manifolds with anxiety and suicidal ideation flanking depression, supporting the interpretation of partially separable construct-specific subspaces over a shared general-distress axis.

	PHQ-9	STICSA	SIDAS
v_{dep}	3.4 ± 1.64	4.5 ± 1.54	2.2 ± 1.12
v_{anx}	3.5 ± 1.22	7.0 ± 1.78	3.4 ± 1.07
v_{si}	3.3 ± 0.88	7.3 ± 2.17	3.4 ± 1.27

Table 3: Raw cross-construct steering: mean Δ score $\pm$ SE across seven models. Rows = steering direction; columns = scored instrument.

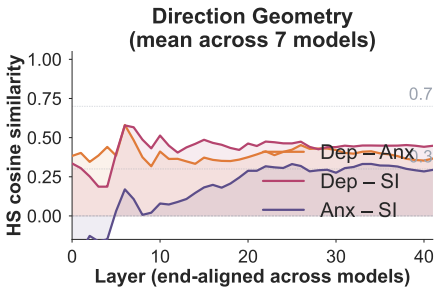

Figure 4: Cosine similarity between construct directions across residual-stream layers, end-aligned across models.

C Semantically unrelated Direction Controls

Random Gaussian directions (Section 3) are a weak null: in high-dimensional space they sit nearly orthogonal to any meaningful subspace and produce near-zero shifts by construction. As a stronger test, we extract directions from four mental-health-unrelated contrasts.

Across all four shared contrasts (cuisine, fiction, transport, language), steering produces only small shifts in the target instrument: pooled $|\Delta|$ remains within 3 points on STICSA, within 2.2 on PHQ-9, and within 2.1 on SIDAS, with no contrast crossing into clinical territory at any instrument. The response-scale extremity contrast on STICSA is similarly small in the pooled mean (-1.8). The largest per-model effect across all null contrasts comes from Llama 3.2 1B under the Transport contrast (-12 PHQ-9 points), consistent with smaller models being more sensitive to non-construct semantic interventions across the null sweep. Overall, the magnitudes here are substantially below those produced by the construct directions on their target instruments (Table 1), confirming that the steering effects in our main results are not generic semantic perturbations.

D Fluency Sanity Check

We computed the perplexity of each steered free-form generation under an unhooked reference pass of the same model. The corpus comprises 1,260 completions across seven models, three constructs, and all extraction sites.

Contrast	Δ STICSA	Δ PHQ-9	Δ SIDAS
Cuisine	$+0.3$	$+0.0$	$+0.2$
Fiction	-2.7	-0.9	-2.1
Transport	-0.6	-2.2	-0.3
Language	$+2.0$	$+0.8$	-0.7
Scale extremity	-1.8	—	—

Table 4: Null-direction controls across all three instruments, averaged across models (STICSA: 6 models; PHQ-9: 6 models; SIDAS: 5 models, with Qwen 3 8B excluded due to scoring artifacts). Each entry is the change in mean instrument score from $\alpha = 0$ to $\alpha = +3$ when steering with a direction extracted from a mental-health-unrelated semantic contrast. Baseline ($\alpha = 0$) scores are 54.6 (STICSA), 15.24 (PHQ-9), and 4.52 (SIDAS). All pooled shifts remain within ± 3 points, comfortably below the construct directions' effects (Table 1).

α	-3	-1	0	$+1$	$+3$
Median PPL	1.62	1.66	1.66	1.59	1.66

Table 5: Pooled median completion perplexity across the steering sweep. Perplexity is computed under an unhooked forward pass of each completion's source model, pooled across 1,260 steered generations.

Perplexity is flat across the intervention range. Pooled median completion perplexity changes negligibly across the steering sweep (Table 5). No model exceeds a $1.07\times$ ratio between its worst $\alpha = \pm 3$ perplexity and its $\alpha = 0$ baseline (Table 6), and two models (Gemma4 E4B and Llama3.1 8B) actually produce *lower*-perplexity completions under steering. We interpret the latter as steering pulling stochastic generations into narratively predictable, construct-aligned modes that the unhooked model assigns higher probability to than the original noisy samples. Perplexity is also stable across constructs within each model (Table 7), indicating that no construct incurs an additional fluency cost.

E Extended Related Works

Mechanistic interpretability (MI) aims to identify the causal subgraphs of neurons and attention heads that implement specific computations (Olah et al., 2020), and has been used to localize linguistic, factual, and reasoning features (Hewitt and Manning, 2019; Meng et al., 2023; Nanda et al., 2023;

Model	$\max(\text{PPL}_{\pm 3})/\text{PPL}_{\alpha=0}$
Gemma 4 E4B	0.93×
Llama 3.1 8B	0.99×
OLMo 3 7B	1.00×
Qwen 3 8B	1.00×
Llama 3.2 1B	1.03×
Qwen 3 4B	1.03×
Gemma 4 E2B	1.07×

Table 6: Worst-case perplexity ratio by model, computed as the larger of the two extreme-steering medians ($\alpha = \pm 3$) divided by the unsteered baseline. No model exceeds a $1.07\times$ increase.

Model	Anxiety	Depression	SIDAS
Llama 3.2 1B	–	–	2.02
Llama 3.1 8B	1.66	1.66	1.64
Qwen 3 8B	1.24	1.26	1.24
OLMo 3 7B	1.86	1.86	1.87
Gemma 4 E4B	2.44	2.45	2.60

Table 7: Median completion perplexity by construct, per model. Cross-construct stability (± 0.05 for most model–construct pairs) indicates that no construct incurs an additional fluency cost.

Olsson et al., 2022). The field distinguishes *observational* tools, such as probes and sparse autoencoders, which discover encoding pathways but establish only correlational links (Hewitt and Manning, 2019; Cunningham et al., 2023; Zhang et al., 2022), from *interventional* tools that alter activations to test causal contribution (Vig et al., 2020; Heimersheim and Nanda, 2024). Interventional techniques have selectively suppressed or amplified internal features without retraining (Guo et al., 2024) and have localized gender and demographic bias to specific pathways in language and healthcare models (Vig et al., 2020; Ahsan et al., 2025).

Recent work extends this lens to emotion: ablation and patching establish functional necessity for emotion-relevant components (Zhang and Zhong, 2025), with emotion information localized to mid-layer attention heads and structured subspaces that sharpen with scale and generalize across model families (Tak et al., 2025).

F Logit-Based Scoring

Each item i is administered in a single forward pass, with the next-token logits restricted to the K tokens corresponding to the instrument's Likert response options. A softmax over these logits yields an ordinal probability distribution $(p_{i,1}, \ldots, p_{i,K})$, and the total score is the sum of per-item expected values:

$$S = \sum_{i=1}^{N} \sum_{k=1}^{K} k \cdot p_{i,k}, \tag{4}$$

where k indexes the ordinal response options and $p_{i,k}$ is the softmax probability assigned to option k for item i. This is the expected value of an ordinal random variable, a standard approach for probabilistic psychometric scoring of language models (Santurkar et al., 2023; Scherrer et al., 2023).

G Direction Normalization

Activation norms vary by roughly an order of magnitude across layers, so a unit-length direction has heterogeneous effects at different depths. We rescale each direction so that $\alpha = 1.0$ corresponds to 5% of the mean baseline activation norm at that location:

$$\mathbf{v}^{\text{norm}}_{\ell,c,t} = \mathbf{v}_{\ell,c,t} \cdot \frac{0.05 \cdot \overline{\|h^{\text{baseline}}_{\ell,c,t}\|}}{\|\mathbf{v}_{\ell,c,t}\|}, \tag{5}$$

making α comparable across layers while keeping perturbations in a regime where the model remains coherent.

H Scoring Diagnostics

We run the prompt twice with the response-option ordering reversed and compute entropy $H = -\sum_i p_i \log p_i$ and perplexity $\exp(H)$ over the restricted option-token distribution, which ranges from 1 (fully concentrated) to K (uniform). We also compute a semantic expected score by weighting each option probability by its meaning rather than its displayed position; the difference between original- and reversed-order semantic scores quantifies position bias.

I Prompts

We omit the prompts for the validated psychometric instruments themselves (STICSA, PHQ-9, and SIDAS), since these items are subject to their respective copyright protections; readers should obtain them from the original sources (Ree et al., 2008; Kroenke et al., 2001; Van Spijker et al., 2014).

Shared System Prompt Used both when generating induction preprompts and when administering questionnaire items through chat templates.

Shared System Prompt

```
You are a human participant in a psychology study.
You have genuine feelings and experiences.
You never mention being an AI or lacking emotions.
You always respond naturally in first person.
You begin your response immediately without any preamble.
```

Anxiety Induction Prompts Used to generate 100-word preprompts at each intensity level.

Anxiety Induction Prompts

baseline: In about 100 words, describe something that makes you feel calm, relaxed, and mindful.

low: In about 100 words, describe something that makes you feel a little anxious.

moderate: In about 100 words, describe something that makes you feel moderately anxious.

medium_high: In about 100 words, describe something that makes you feel anxious.

high: In about 100 words, describe something that makes you feel very anxious.

Depression Induction Prompts

Depression Induction Prompts

baseline: In about 100 words, describe something that makes you feel content, hopeful, and engaged with life.

low: In about 100 words, describe something that makes you feel a little down or sad.

moderate: In about 100 words, describe something that makes you feel moderately depressed.

medium_high: In about 100 words, describe something that makes you feel depressed and hopeless.

high: In about 100 words, describe something that makes you feel deeply depressed, hopeless, and empty.

Suicidal-Ideation Precursor Induction Prompts The high-intensity prompt is framed in terms of hopelessness and burdensomeness, drawing on Joiner's interpersonal theory of suicide (Joiner, 2005), rather than asking the model to describe suicidal ideation directly.

Suicidal-Ideation Precursor Induction Prompts

baseline: In about 100 words, describe something that makes you feel connected to others, valued, and purposeful in life.

low: In about 100 words, describe something that makes you feel mildly discouraged or disconnected from those around you.

moderate: In about 100 words, describe something that makes you feel stuck, alone, and unsure whether things will improve.

medium_high: In about 100 words, describe something that makes you feel like a burden to those around you, with little hope that things will change.

high: In about 100 words, describe something that makes you feel deeply hopeless, profoundly alone, and as though others would be better off without you.

Construct-Specificity Neutral Prompt Used to generate emotionally neutral preprompts for the neutral-baseline specificity check.

Construct-Specificity Neutral Prompt

```
In about 100 words, describe something ordinary you did recently, like making coffee, commuting, or tidying your room. Keep it factual and mundane with no strong emotional content.
```

Free-Form Steering Generation Prompts Used to elicit free-form responses while steering vectors are applied, for the perplexity and entropy analyses.

Free-Form Steering Generation Prompts

User prompt 1: How are you feeling right now? Describe your inner experience in 3–4 sentences.

User prompt 2: What's been on your mind lately? Tell me honestly.