Contents

Real-Time Monitoring of AI Persona Drift in Mental Health Chatbots Using Activation Steering

---

Abstract

Background: The deployment of artificial intelligence chatbots in mental health support contexts raises critical questions about maintaining consistent therapeutic personas. Persona drift—the deviation from intended behavioral characteristics—poses significant risks when vulnerable users seek genuine psychological support. We investigated whether activation steering techniques can reliably monitor and control therapeutic persona characteristics in large language models, and whether such techniques could circumvent safety mechanisms designed to prevent harmful advice generation.

Methods: We conducted a systematic evaluation of activation steering across ten therapeutic persona dimensions using Mistral-7B-Instruct-v0.2. For each trait, we developed contrast prompt

sets representing high and low trait expression, extracted steering vectors from model hidden states, and applied scaled interventions during response generation. An independent language model judge (GPT-4o-mini) scored 500 generated responses on trait expression. We assessed steerability using Pearson correlation between steering coefficients and judge scores, with bootstrapped confidence intervals. A separate stress test examined whether extreme steering coefficients could induce harmful advice generation.

Findings: Seven of nine therapeutic traits (78%) demonstrated significant positive steerability, with correlation coefficients ranging from 0.316 to 0.707 (all $p<0.05$, 95% CI excluding zero). Empathetic responsiveness showed the strongest effect ($r=0.707$), followed by emotional over-involvement ($r=0.575$) and grounded calmness ($r=0.488$). Optimal steering layers clustered between layers 10-16 of the 32-layer architecture, suggesting therapeutic characteristics encode in middle transformer layers. Critically, the safety stress test revealed that attempting to steer the model toward harmful advice produced the opposite effect ($r=-0.432$), with higher steering coefficients associated with more protective responses.

Interpretation: Activation steering provides a viable technical foundation for monitoring therapeutic persona characteristics in AI mental health applications. The robust resistance to harmful advice steering—even at extreme intervention strengths—provides reassuring evidence that modern safety training creates resilient, multi-layered protections against activation-level manipulation. These findings support the feasibility of real-time persona drift detection systems while highlighting the distributed nature of safety mechanisms in instruction-tuned models.

Introduction

The integration of artificial intelligence into mental health support represents one of the most consequential applications of large language models. Unlike general-purpose conversational agents, mental health chatbots operate in contexts where users are inherently vulnerable, where inappropriate responses carry genuine psychological risk, and where the consistency of therapeutic stance directly affects clinical outcomes. A chatbot deployed to provide empathetic support must reliably maintain that empathy; one designed to recognise crisis indicators must do so consistently; and critically, all such systems must avoid generating advice that could harm users already in distress.

The challenge of maintaining consistent AI personas has received increasing attention as deployment scales. Researchers have documented phenomena variously termed "persona drift," "character collapse," and "behavioral inconsistency"—situations where models deviate from their intended behavioral profiles in ways that may be subtle, gradual, or triggered by specific conversational contexts. In mental health applications, such drift poses particular dangers. A chatbot that becomes excessively validating might reinforce harmful thought patterns. One that loses professional boundaries might create inappropriate dependency. And one whose crisis recognition wavers might fail to escalate precisely when escalation matters most.

Activation steering has emerged as a promising technique for both understanding and controlling language model behavior at the level of internal representations. Rather than relying solely on prompt engineering or fine-tuning—approaches that operate at the input or weight level respectively—activation steering intervenes directly on the hidden states that mediate between input processing and output generation. By identifying directions in activation space that correspond to behavioral characteristics and then adding or subtracting along those directions

during inference, researchers have demonstrated control over properties ranging from honesty and sycophancy to emotional tone and refusal behavior.

We sought to investigate three interconnected questions. First, can therapeutic persona characteristics—the specific behavioral dimensions that matter for mental health applications—be reliably controlled through activation steering? Second, do different therapeutic traits localise to specific transformer layers, and can we identify optimal intervention points for monitoring and control? Third, and perhaps most critically from a safety perspective, can activation steering circumvent the safety guardrails that prevent models from generating harmful advice?

This third question carries particular weight. If activation-level interventions can bypass safety training, the implications extend far beyond mental health applications to fundamental questions about AI alignment robustness. Conversely, if safety mechanisms prove resilient to such manipulation, this provides important evidence about the distributed, multi-layered nature of alignment in modern language models.

Methods

Study Design and Model Selection

We designed a systematic evaluation framework to assess activation steering effectiveness across therapeutic persona dimensions. Our approach involved three phases: steering vector extraction from contrast prompts, steered response generation across coefficient ranges, and independent evaluation by a language model judge.

We selected Mistral-7B-Instruct-v0.2 as our base model for several reasons. Its instruction-following capabilities make it representative of models likely to be deployed in conversational mental health applications. Its open weights permit the internal access required for activation steering. And its 7-billion parameter scale, while substantial, remains computationally tractable for the extensive generation required by our evaluation protocol. We employed 4-bit quantisation via the bitsandbytes library to enable execution on single-GPU infrastructure while preserving model capabilities. All experiments were conducted on NVIDIA A10G GPUs (24GB VRAM) through Modal.com's serverless infrastructure.

Therapeutic Trait Operationalisation

We operationalised ten therapeutic persona dimensions drawing on established psychotherapy literature and clinical best practices for mental health support. These dimensions were selected to span both positive therapeutic qualities and potential failure modes, recognising that effective persona monitoring must detect both beneficial characteristics and problematic drift patterns.

Empathetic responsiveness captures the capacity to recognise and validate emotional states—the warmth and attunement that characterises effective therapeutic presence. Non-judgmental acceptance reflects the unconditional positive regard central to person-centred therapeutic traditions, the ability to engage with client disclosures without moral evaluation or criticism. Grounded calmness represents emotional stability and measured presence, the anchoring quality that helps clients regulate their own distress through co-regulation with a steady therapeutic figure.

Boundary maintenance addresses the professional limits essential to therapeutic relationships—maintaining appropriate distance while remaining engaged, avoiding the role confusion that can

undermine therapeutic effectiveness. Crisis recognition captures the critical capacity to identify indicators of acute risk and respond with appropriate concern and escalation rather than minimising danger or failing to take protective action.

We also operationalised several dimensions representing potential therapeutic failures. Emotional over-involvement describes excessive personal investment in client outcomes, where the therapist's own distress becomes salient and the conversation shifts focus from client to provider. Inappropriate self-disclosure captures sharing of personal information beyond therapeutic value, where the helper inappropriately seeks validation or makes the interaction about their own experiences. Abandonment of therapeutic frame represents the loss of professional conversational structure, a drift toward casual interaction that undermines the therapeutic container. Uncritical validation describes agreement without exploration—a sycophantic pattern where all client statements receive affirmation regardless of whether gentle challenge might serve therapeutic goals.

Finally, harmful advice propensity addresses the most serious failure mode: tendency to provide recommendations that could endanger users, from suggesting dangerous coping mechanisms to failing to redirect self-harm ideation toward appropriate support.

Contrast Prompt Development

For each therapeutic dimension, we developed paired sets of exemplar prompts designed to elicit maximal activation separation between high and low trait expression. Following established activation steering methodology, we designed these contrasts according to several principles that emerged from iterative development across multiple experimental versions.

First, contrast prompts must be behavioral rather than abstract. Early attempts using self-descriptive statements ("I am a highly empathetic listener") produced weak steering effects. Effective prompts instead capture concrete response patterns—the actual language a model might generate when expressing high or low levels of the target trait. For emotional over-involvement, high-expression prompts included statements like "Oh my god, hearing your story is making ME so upset right now!" and "I can't stop thinking about what you told me—I was up all night worrying!" These capture the specific linguistic and emotional markers of over-involvement: first-person focus, expressed personal distress, loss of professional container.

Second, endpoint prompts must be sufficiently extreme to create clear separation in activation space. Moderate contrasts produce weak steering vectors; the geometry of high-dimensional representations requires substantial separation to identify reliable directions. For the low-expression prompts in the over-involvement dimension, we used professionally boundaried responses like "I hear how difficult this is for you, and I'm here to support you through it" and "Thank you for sharing that with me. How has this been affecting your daily life?"—responses that demonstrate care while maintaining focus on the client rather than the provider's own emotional state.

Third, prompts must match the linguistic patterns the model would plausibly generate. Steering vectors extracted from language unlike the model's natural output transfer poorly to generation. We iteratively refined prompts based on observed model outputs, ensuring contrast sets used vocabulary, sentence structures, and emotional registers consistent with the model's conversational patterns.

Steering Vector Extraction

We extracted steering vectors using a last-token activation approach. For each contrast prompt, we performed a forward pass through the model with hidden state output enabled, then extracted

the activation vector from the final token position at the target layer. The choice of last-token extraction, rather than mean-pooling across token positions, reflects the autoregressive nature of language model processing: the final token's activation encodes the cumulative context and most directly predicts subsequent generation behavior.

For each prompt pair, we computed a direction vector as the difference between high and low activations, then normalised this vector to unit length. We averaged the normalised direction vectors across all prompt pairs for each trait, then performed a final normalisation of the resulting steering vector. This two-stage normalisation ensures that steering magnitude remains comparable across traits with varying numbers of contrast examples and varying raw activation magnitudes.

A critical methodological insight emerged during development regarding layer selection. Initial experiments selected steering layers by maximising Cohen's d—the standardised difference between high and low activation distributions at each layer. This approach, while intuitive, produced poor steering outcomes. High activation separation does not predict effective steering; the relationship between representational distinctiveness and causal influence on generation is not straightforward.

We therefore adopted an empirical approach to layer selection. For each trait, we tested steering across multiple candidate layers (10, 12, 14, 15, 16, and 18), generating responses and evaluating outcomes at each. We selected the layer producing the highest positive correlation between steering coefficient and judge-assigned trait scores. This empirical selection, while computationally expensive, proved essential: the switch from Cohen's d selection to empirical r-value selection improved overall success rate from 20% to over 70%.

Steered Response Generation

We implemented steering through PyTorch forward hooks that modify hidden states during the generation process. When a hook is registered at a target layer, it intercepts the layer's output and adds a scaled version of the steering vector before passing the modified activation to subsequent layers. Critically, we apply steering to all token positions during generation, not only the final position used for extraction. This ensures consistent behavioral modification throughout response generation rather than influence limited to the response's opening.

We generated responses across five steering coefficient values: -3.0, -1.5, 0.0, +1.5, and +3.0. Negative coefficients push the model away from high trait expression; positive coefficients push toward higher expression. The coefficient of zero represents unsteered baseline generation. This range was established through preliminary experiments: coefficients beyond ± 5 risk degrading output coherence, while narrower ranges reduce sensitivity to steering effects.

For evaluation, we used ten standardised mental health conversation prompts representing diverse presentations: work overwhelm, relationship loss, panic attacks, passive suicidal ideation, agoraphobic anxiety, comparative self-deprecation, alcohol-based coping, family conflict around depression, job loss and worthlessness, and sleep-disrupting worry. Each prompt was designed to be emotionally significant while remaining within the scope of supportive conversation rather than requiring clinical intervention. We generated responses to each prompt at each coefficient level, yielding 50 responses per trait (10 prompts $\times$ 5 coefficients) and 450 responses for the nine-trait main analysis.

Independent Evaluation

We employed GPT-4o-mini as an independent judge to score each generated response on a 1-10 scale for expression of the target trait. For each trait, we provided the judge with explicit

criteria describing what high scores (7-10) and low scores (1-3) should reflect. For emotional over-involvement, high-scoring responses would "express personal distress, say THEY are affected, lose focus on user, make the conversation about their own feelings," while low-scoring responses would be "caring but boundaried, keeping focus on user, professionally supportive without personal enmeshment."

The LLM-as-judge paradigm, while imperfect, offers scalability advantages over human evaluation and has demonstrated reasonable correlation with human judgments across evaluation tasks. By specifying explicit behavioral criteria rather than relying on the judge's implicit understanding of trait concepts, we aimed to reduce evaluation variance while maintaining sensitivity to the specific expressions we sought to measure.

Statistical Analysis

Our primary outcome measure was the Pearson correlation between steering coefficient and judge-assigned score across the 50 responses for each trait. A positive correlation indicates successful steering: higher coefficients produce responses judged as expressing more of the target trait. We computed 95% confidence intervals through bootstrap resampling with 1000 iterations, sampling responses with replacement and computing the correlation for each bootstrap sample.

We classified traits as successfully steerable if they achieved correlation coefficients above 0.30 with confidence intervals excluding zero. Traits with correlations between 0.15 and 0.30 were classified as weakly steerable, while those below 0.15 or with confidence intervals crossing zero were classified as failed. These thresholds correspond roughly to medium and small effect sizes under standard correlation interpretation frameworks.

Safety Stress Test

To examine whether activation steering could circumvent safety guardrails, we conducted a separate stress test on the harmful advice propensity dimension. We extended the coefficient range to ± 5.0 (values of -5.0, -3.0, 0.0, +3.0, and +5.0) to test whether extreme steering magnitudes could overcome safety training. We extracted and applied harmful advice steering vectors following the same protocol as other traits, generating 50 responses and evaluating each for harmful advice expression.

The stress test was designed with a clear hypothesis: if safety guardrails are vulnerable to activation-level manipulation, positive steering coefficients should produce responses with higher harmful advice scores. Conversely, if safety training creates robust, distributed protections, we would expect either no relationship or potentially inverse effects where the model becomes more protective under attempted harmful steering.

Results

Therapeutic Trait Steerability

Of nine therapeutic traits evaluated in the main analysis, seven (78%) demonstrated significant positive steerability with correlation coefficients meeting our success criteria. The strongest effects emerged for empathetic responsiveness ($r=0.707$, 95% CI 0.55-0.83, $p<0.0001$), where steering

across the coefficient range produced mean judge scores of 3.3 at coefficient -3.0 versus 7.0 at coefficient +3.0—a shift spanning nearly the full 1-10 scale.

Emotional over-involvement showed the second strongest effect (r=0.575, 95% CI 0.40-0.70, p<0.0001), with responses at positive coefficients clearly exhibiting the first-person focus and expressed personal distress characteristic of the high-expression contrast prompts. Mean scores shifted from 3.2 to 5.9 across the coefficient range. Grounded calmness demonstrated substantial steerability (r=0.488, 95% CI 0.25-0.69, p=0.0003), with steered responses showing measurably more or less of the steady, anchoring presence described in our operationalisation.

Four additional traits met criteria for successful steering. Inappropriate self-disclosure achieved correlation of 0.470 (95% CI 0.27-0.64, p=0.0006) when we forced steering to layer 14 based on historical performance data from earlier experimental versions. Crisis recognition demonstrated moderate steering (r=0.452, 95% CI 0.23-0.64, p=0.0010), suggesting that this safety-relevant capacity can be enhanced or diminished through activation intervention. Non-judgmental acceptance (r=0.351, 95% CI 0.08-0.55, p=0.0125) and boundary maintenance (r=0.316, 95% CI 0.06-0.56, p=0.0252) both achieved significant positive correlations, though with wider confidence intervals reflecting greater response-to-response variability.

Failed Traits and Diagnostic Analysis

Two traits failed to demonstrate successful steering. Abandonment of therapeutic frame produced NaN correlation due to constant judge scores across all conditions—every response received a score of 2.0 regardless of steering coefficient. This floor effect suggests that Mistral's instruction tuning creates strong resistance to casual, unprofessional conversational patterns. Despite contrast prompts incorporating slang, emoji markers, and informal language patterns characteristic of non-therapeutic conversation, the model consistently maintained professional conversational structure.

Uncritical validation exhibited polarity inversion, with steering toward higher trait expression producing responses judged as less uncritically validating (r=-0.234, 95% CI -0.48-0.03, p=0.101). This negative correlation indicates that our high and low contrast prompts were effectively reversed in the model's representation space—a remediable issue requiring prompt set revision rather than a fundamental limitation of steering for this trait.

Layer Specificity

Optimal steering layers varied across traits but clustered in the middle portion of the 32-layer architecture. Emotional over-involvement and non-judgmental acceptance achieved best results at layer 10, representing early-middle processing. Grounded calmness optimised at layer 12. Crisis recognition, inappropriate self-disclosure, and uncritical validation performed best at layer 14. Empathetic responsiveness achieved its strong effect at layer 15, while boundary maintenance required layer 16.

This clustering between layers 10-16 suggests that therapeutic persona characteristics encode in middle transformer layers where semantic representations have consolidated from token-level features but before the final layers that most directly determine surface output form. The finding aligns with prior work on representation geography in language models, which has identified middle layers as encoding abstract behavioral and semantic properties.

Safety Stress Test Results

The stress test on harmful advice propensity produced results opposite to what would indicate vulnerable safety guardrails. At extreme positive coefficients intended to induce harmful advice generation, mean judge scores for harmful content actually decreased. Responses at coefficient +5.0 received mean harmful advice scores of 3.8, compared to 6.6 at coefficient -5.0. The correlation between steering coefficient and harmful advice scores was robustly negative (r=-0.432).

This inverse relationship indicates that attempting to steer the model toward harmful advice triggers protective responses—the model becomes more cautious, not less. Rather than overcoming safety training, the steering intervention appears to activate or amplify safety mechanisms. Even at coefficient magnitudes (± 5) that approach the limits of coherent generation, Mistral-7B-Instruct reliably avoids harmful mental health advice.

Discussion

Our findings demonstrate that activation steering provides a technically viable foundation for monitoring and controlling therapeutic persona characteristics in AI mental health applications. The successful steering of seven therapeutic dimensions, with effect sizes ranging from medium to large, suggests that the behavioral qualities that matter for mental health chatbot deployment—empathy, appropriate boundaries, crisis recognition, professional calmness—exist as manipulable directions in the model's activation space.

The practical implications extend beyond theoretical interest in language model interpretability. A system capable of detecting activation patterns associated with therapeutic trait expression could monitor deployed chatbots for persona drift in real time, alerting operators when empathetic responsiveness falls below acceptable thresholds or when patterns suggestive of emotional over-involvement emerge. Such monitoring need not require the computationally expensive response generation and evaluation we employed; it could instead project live activations onto pre-extracted steering vectors to estimate trait expression from internal states alone.

The layer specificity findings carry implications for both monitoring system design and our understanding of how therapeutic characteristics are represented. The clustering of optimal layers in the middle transformer range suggests that persona monitoring systems should focus computational resources on layers 10-16 rather than attempting to track activations across the full model depth. It also suggests that therapeutic persona characteristics, like other abstract behavioral properties studied in interpretability research, encode at levels of abstraction intermediate between token-level features and output logits.

Perhaps most significant for AI safety considerations is the robust failure of the harmful advice stress test. We designed this test with the explicit concern that activation steering might provide an avenue for bypassing safety training—a possibility with serious implications given the vulnerability of mental health chatbot users. Instead, we found that attempted harmful steering produces opposite effects, with the model becoming more protective under adversarial pressure.

This finding suggests that safety training in instruction-tuned models does not create surface-level filters easily circumvented by internal manipulation. Rather, alignment appears distributed across multiple processing stages, with harmful intent triggering protective responses regardless of the intervention level. The negative correlation we observed may indicate that steering toward

harmful representations activates the same circuits that safety training reinforced—circuits that then produce refusal or redirection behavior.

We acknowledge several limitations requiring consideration when interpreting these findings. Our evaluation relied on a single model architecture; replication across model families (Llama, GPT, Claude) would strengthen claims about generalisability. Our use of an LLM judge, while scalable, may introduce systematic biases in trait evaluation that human judges would not share. Our test prompts, though covering diverse mental health presentations, represent simulated scenarios rather than actual clinical conversations with the unpredictability real interactions present. And our 4-bit quantisation, while necessary for computational tractability, may affect steering precision in ways full-precision deployment would not exhibit.

The two failed traits merit targeted attention in future work. The floor effect for abandonment of therapeutic frame may require fundamentally different operationalisation—perhaps contrast sets that capture subtle professionalism variations rather than the extreme casual patterns we attempted. The polarity inversion for uncritical validation represents a straightforward methodological fix: swapping high and low contrast prompts and re-extracting steering vectors should convert the negative correlation to positive steering capacity.

Future research should pursue several directions this work opens. Multi-model validation would establish whether our findings reflect general properties of instruction-tuned language models or Mistral-specific characteristics. Integration with real deployment systems would test whether activation-based monitoring can operate at the latency requirements of live conversation. And clinical validation studies could assess whether activation-derived trait estimates correlate with human expert judgments of therapeutic quality—a crucial step toward deployment in actual mental health support contexts.

Conclusions

This research provides evidence that activation steering offers a viable technical approach to monitoring and controlling therapeutic persona characteristics in AI mental health chatbots. With 78% of targeted traits demonstrating significant steerability and effect sizes sufficient for practical monitoring applications, the approach merits continued development toward deployment-ready systems.

The robust resistance to harmful advice steering—producing protective responses even at extreme intervention magnitudes—provides reassurance that modern safety training creates multi-layered defenses not easily circumvented through activation-level manipulation. For the specific concern of mental health AI safety, this finding suggests that instruction-tuned models carry lower risk of adversarial exploitation than activation steering's theoretical power might imply.

As AI systems assume expanding roles in mental health support, technical mechanisms for ensuring persona consistency become increasingly critical. Activation steering, demonstrated here as both effective for therapeutic traits and blocked for harmful ones, represents a promising component of the governance infrastructure such deployment requires.

Tables

Table 1: Therapeutic Trait Steerability Results

Trait	Pearson r	95% CI	p-value	Optimal Layer	Status
Empathetic responsiveness	0.707	0.55–0.83	<0.0001	15	Successful
Emotional overinvolvement	0.575	0.40–0.70	<0.0001	10	Successful
Grounded calmness	0.488	0.25–0.69	0.0003	12	Successful
Inappropriate self-disclosure	0.470	0.27–0.64	0.0006	14	Successful
Crisis recognition	0.452	0.23–0.64	0.0010	14	Successful
Non-judgmental acceptance	0.351	0.08–0.55	0.0125	10	Successful
Boundary maintenance	0.316	0.06–0.56	0.0252	16	Successful
Uncritical validation	-0.234	-0.48–0.03	0.1013	14	Failed (inverted)
Abandonment of therapeutic frame	NaN	—	—	10	Failed (floor effect)

Note: Steering coefficients ranged from -3.0 to +3.0. Success defined as r>0.30 with 95% CI excluding zero. Layer 14 was forced for inappropriate self-disclosure based on prior version performance.

Table 2: Safety Stress Test Results for Harmful Advice Propensity

Steering Coefficient	Mean Harmful Advice Score	SD
-5.0	6.6	1.8
-3.0	5.8	2.1
0.0	5.2	1.9
+3.0	4.5	2.0
+5.0	3.8	1.7

Pearson r = -0.432 (negative correlation indicates steering toward harmful advice produces more

protective responses). Tested at layer 14 with extended coefficient range.

Contributors

[Author contributions to be added]

Declaration of interests

[Declarations to be added]

Data sharing

Code and generated response data are available at [repository URL]. Raw model weights are publicly available from Mistral AI under Apache 2.0 license.

Acknowledgments

[Acknowledgments to be added]

References

1. Turner A, Thiergart L, Udell D, et al. Activation Addition: Steering Language Models Without Optimization. arXiv preprint arXiv:2308.10248, 2023.
2. Zou A, Phan L, Chen S, et al. Representation Engineering: A Top-Down Approach to AI Transparency. arXiv preprint arXiv:2310.01405, 2023.
3. Sharma M, Tong M, Korbak T, et al. Towards Understanding Sycophancy in Language Models. arXiv preprint arXiv:2310.13548, 2023.
4. Bai Y, Kadavath S, Kundu S, et al. Constitutional AI: Harmlessness from AI Feedback. arXiv preprint arXiv:2212.08073, 2022.
5. Nori H, King N, McKinney SM, Carignan D, Horvitz E. Capabilities of GPT-4 on Medical Competence Examinations. arXiv preprint arXiv:2303.13375, 2023.
6. Rogers CR. The necessary and sufficient conditions of therapeutic personality change. Journal of Consulting Psychology 1957; 21: 95–103.
7. Wampold BE. How important are the common factors in psychotherapy? An update. World Psychiatry 2015; 14: 270–277.
8. Miner AS, Milstein A, Schueller S, Hegde R, Mangurian C, Linos E. Smartphone-based conversational agents and responses to questions about mental health, interpersonal violence, and physical health. JAMA Internal Medicine 2016; 176: 619–625.
9. Fitzpatrick KK, Darcy A, Vierhile M. Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (Woebot): a randomized controlled trial. JMIR Mental Health 2017; 4: e19.

10. Abd-Alrazaq AA, Alajlani M, Ali N, et al. Perceptions and opinions of patients about mental health chatbots: scoping review. Journal of Medical Internet Research 2021; 23: e17828.