How Do Internal Affective Representations Relate to Self-Report in LLMs?

Anonymous authors
Paper under double-blind review

Abstract

Large language models (LLMs) are increasingly used in affect-sensitive settings such as psychological support, where sensitivity to human affect may need to be preserved while potentially misleading self-referential affective expression is selectively regulated. We investigate whether two such capabilities—predicting the affective response that a text is likely to evoke in a human reader (Reader Prediction) and producing an affective Self-Report for the same text—rely on shared internal representations and causal mechanisms. Using a matched-stimulus design across four model families, we compare the two tasks at the levels of behavioral correspondence, representation, and causal utilization, and examine differences between Base and Instruct models and across generation stages. Reader Prediction and Self-Report show strong behavioral coupling, and affect-relevant information is decodable from both tasks; however, evidence for shared representations and cross-task causal utilization is limited, indicating that behavioral correspondence and decodability do not imply a shared mechanism. Base–Instruct comparisons further reveal differences in representational organization and causal utilization, while layer- and generation-stage analyses provide limited evidence for robust endogenous control of Self-Report by decodable affect-relevant information. Overall, our results distinguish behavioral correspondence, representational sharing, and causal utilization as separate levels of evidence, suggesting that affect-related tasks may share structured information without utilizing it in the same way. This distinction provides a mechanistic basis for selective interventions that preserve human-affect prediction while regulating self-referential affective expression.

1 Introduction

Large language models (LLMs) are increasingly used in affect-sensitive contexts, including psychological support and interpersonal advice (Li et al., 2023a; Luo et al., 2025). In such settings, models may predict the affective response that a situation or text would evoke in a human reader, while also generating self-referential statements about their own apparent affective responses. Such expressions may shape how users perceive and interact with AI systems, particularly when interpreted as evidence that the model itself has emotions (Weidinger et al., 2022; Laestadius et al., 2024; Fang et al., 2025).

This raises a mechanistic question: when an LLM predicts a human reader's affective response and reports its own response to the same stimulus, to what extent does it rely on the same internal representations. Similar outputs need not imply identical mechanisms: the two tasks may share affect-relevant representations while differing in how those representations are used. This distinction matters for selective regulation of affective behavior. In applications such as psychological support, sensitivity to human affect may need to be preserved while limiting potentially misleading self-referential affective expression. If the two capabilities rely on shared mechanisms, suppressing one may disrupt the other; if they are only partially shared, more selective interventions may be possible.

Prior work shows that LLMs can generate responses perceived as empathetic (Ayers et al., 2023; Elyoseph et al., 2023; Huang et al., 2023; Yu et al., 2025) and that sentiment- and emotion-related information can be decoded from their hidden representations (Di Palma et al., 2025; Tak et al., 2025; Wu et al., 2026). However, these findings do not establish the mechanistic relationship between affect prediction and affective Self-Report. Behavioral similarity does not imply shared internal

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

representations, decodability does not imply causal utilization (Hewitt & Liang, 2019), and explicit Self-Report need not faithfully reflect the internal representations involved in generation (Burns et al., 2022; Turpin et al., 2023). We therefore distinguish behavioral correspondence, representational sharing, and causal utilization.

We study this relationship using a matched-stimulus design. In **Reader Prediction**, the model estimates the affective response that a text is likely to evoke in an average human reader. In **Self-Report**, the model is prompted to numerically characterize its response to the same text along affective dimensions. We use *Self-Report* as an operational term for the elicited model output, without assuming subjective emotions, consciousness, or sentience. Applying the two tasks independently to the same stimuli allows us to test whether behavioral correspondence extends to shared representations and causal utilization.

Our central research question is:

To what extent do Reader Prediction and Self-Report rely on shared affect-relevant representations, how are these representations causally utilized across the two tasks, and how do these relationships vary with post-training and across stages of computation?

We operationalize this question through four progressively more mechanistic questions:

1. **Do Reader Prediction and Self-Report covary behaviorally?** We test correspondence across stimuli and responses to controlled affective manipulations.
2. **Do the two tasks share affect-relevant representations and causal mechanisms?** We test representational sharing and cross-task causal utilization.
3. **How does this relationship differ between Base and Instruct models?** We examine differences in representation, cross-task sharing, causal utilization, and output readout.
4. **Where and when is affect-relevant information causally utilized in Self-Report?** We trace causal effects across layers and generation stages, distinguishing decodability from endogenous causal relevance.

Our main contributions are:

1. We introduce a matched-stimulus framework for distinguishing behavioral correspondence between Reader Prediction and Self-Report from mechanistic sharing.
2. We combine representational analyses, semantic controls, and causal interventions to distinguish decodability, representational sharing, and causal utilization.
3. We characterize how these relationships differ between Base and Instruct models and across layers and generation stages.

2 Related Work

Empathy and Affective Behavior in LLMs. A growing body of behavioral work has shown that LLMs can generate responses that humans perceive as empathetic or affectively appropriate. Studies comparing ChatGPT with physicians' responses (Ayers et al., 2023), evaluating emotional awareness (Elyoseph et al., 2023), and testing models on emotion-eliciting situations (Huang et al., 2023) have demonstrated a range of affect-related behaviors. These studies, however, primarily characterize models at the level of their outputs, leaving open whether different affective behaviors—such as predicting a human reader's affective response and producing an affective Self-Report—depend on shared internal mechanisms.

Affect-Relevant Internal Representations and Causal Intervention. A complementary line of work has examined how affective information is represented within LLMs. Sentiment- and emotion-related information can be linearly decoded from hidden states (Di Palma et al., 2025), while interventions along emotion-related activation directions can systematically alter model outputs (Tak et al., 2025; Banayeeanzade et al., 2026). Recent studies have further characterized the geometry of emotion-related representations and their organization across layers and token positions (Zhang & Zhong, 2025; Reichman et al., 2026; Wu et al., 2026). Together, these findings provide evidence

that affect-related information is represented internally and that interventions on affect-related activations can causally alter model outputs. What remains less clear is whether the same affect-relevant representations are shared and causally utilized across distinct affective tasks.

Internal Information and Self-Report. A separate literature has highlighted that information represented internally need not be faithfully reflected in a model's explicit reports. Work on latent knowledge (Burns et al., 2022) and faithfulness (Turpin et al., 2023) demonstrates potential dissociations between internal information and surface-level outputs. More recently, studies of model introspection have investigated whether models can report aspects of their own internal states or predict their own behavior (Binder et al., 2025; Lindsey, 2026; Martorell & Bianchi, 2026; Singh et al., 2026). We do not treat affective Self-Report as evidence of subjective experience; instead, we study its mechanistic relationship to affect-relevant representations within the model.

Post-Training and Internal Mechanisms. Post-training can substantially alter model behavior without necessarily implying the acquisition or removal of the corresponding internal information. Behavioral differences may instead arise from changes in how existing representations are organized, utilized, or mapped onto output policies (Sharma et al., 2024; Ferrando et al., 2025). This distinction is particularly relevant when comparing Base and Instruct models: differences in affective behavior may reflect changes in representation, causal utilization, readout, or some combination of these factors.

Taken together, these lines of work have largely examined affective behavior, internal affect representations, self-report, and post-training as separate questions. Their intersection—how affective Self-Report relates to the representations and causal mechanisms involved in predicting human affect—remains comparatively underexplored.

3 Experimental Setup

Our experiments use a matched-stimulus design in which Reader Prediction and Self-Report are independently elicited for the same texts. We first describe the common tasks, datasets, models, and output measurement procedure, and then summarize the four experimental stages in Table 1. Full experimental specifications, prompts, controls, statistical procedures, and model-family-specific details are provided in the corresponding appendices.

3.1 Tasks and Matched-Stimulus Design

We apply two tasks to the same textual stimuli. In Reader Prediction, the model estimates the affective response that a text is likely to evoke in an average human reader. In Self-Report, the model is prompted to numerically characterize its response to the same text along affective dimensions. We use Self-Report as an operational label for this elicited output, without assuming subjective emotional experience.

Crucially, the two tasks are elicited independently: they use separate prompts and forward passes, and neither task receives the output of the other. This design allows us to compare Reader Prediction and Self-Report without direct information transfer between them.

3.2 Datasets

We use two complementary datasets: EmoBank and AIPsy-Affect.

EmoBank (Buechel & Hahn, 2017) provides continuous human Valence–Arousal–Dominance (VAD) annotations from Reader and Writer perspectives. We use these annotations as an external human reference for evaluating model outputs and affect-relevant hidden representations.

AIPsy-Affect (Keeman, 2026; Labs, 2026) provides stimuli designed for controlled manipulation of affective content. It contains Clinical, Neutral, Moderate, and Complex Neutral stimuli, including 192 matched Clinical–Neutral pairs and 48 matched triplets. These matched stimuli allow us to examine how controlled changes in stimulus content affect model behavior and internal representations. Because AIPsy-Affect does not provide stimulus-level continuous human Valence/Arousal annotations, we do not treat its stimuli as having human V/A ground truth.

Table 1: Overview of the four-stage experimental framework.

Stage	Data	Primary Comparison	Question	Details
Behavioral	EmoBank, AIPsy	Reader ↔ Self	Do Reader Prediction and Self-Report correspond behaviorally and respond similarly to controlled stimulus changes?	App. D
V1	EmoBank, AIPsy	Reader ↔ Self within model	Do the two tasks share affect-relevant representations, and are these representations causally used across tasks?	App. H
V2	EmoBank	Base ↔ Instruct	How do affect-relevant representations and their utilization differ between Base and Instruct models?	App. L
V3	AIPsy	Layer × generation stage	Where and when is affect-relevant information causally utilized during generation?	App. P

3.3 Models

We study four model families: Qwen 2.5 (Qwen et al., 2025), Llama 3.2 (Grattafiori et al., 2024), Gemma 3 (Team et al., 2025), and OLMo 2 (OLMo et al., 2024b;a). The Behavioral, V1, and V2 stages include both Base and Instruct variants, yielding eight models in total, whereas V3 uses the four Instruct models. Full model specifications are provided in Appendix C.

3.4 Output Measurement

For both tasks, we characterize the model's distribution over possible affective reports rather than treating a single greedy generation as the measurement. We define a finite candidate space of valid JSON responses, compute the conditional sequence log-likelihood of each candidate, and Softmax-normalize these values over the candidate space. Expected Valence, Arousal, and related scores are then computed from the resulting distribution.

The Behavioral stage and V1 use a 729-state VAD candidate space, whereas V2 and V3 use an 81-state VA candidate space. Because these candidate spaces define different probability spaces, absolute expected scores are not directly comparable across stages. Cross-stage interpretation therefore focuses on within-stage contrasts and on representational and causal quantities that do not require direct comparison of these expected scores.

3.5 Four-Stage Experimental Framework

Table 1 summarizes the four experimental stages and the primary question addressed by each.

Behavioral Stage: Output-Level Coupling. The Behavioral stage establishes the output-level relationship between Reader Prediction and Self-Report. Using EmoBank, we evaluate how each task corresponds to human affect annotations and how strongly the two model outputs covary across stimuli. Using the matched stimuli in AIPsy-Affect, we further test whether controlled changes in stimulus content induce corresponding changes in Reader Prediction and Self-Report. This stage establishes behavioral coupling but does not by itself imply shared internal mechanisms.

V1: Shared Representation and Causal Utilization. V1 moves from behavioral correspondence to internal mechanism. Within each model, we examine whether affect-relevant information is decodable from hidden states, whether Reader Prediction and Self-Report exhibit shared representational geometry, and whether their apparent sharing remains after controlling for semantic information. We then use causal interventions to test whether affect-relevant information identified in one task can be causally transferred across tasks. Together, these analyses distinguish the presence of affect-relevant information from its shared causal utilization.

V2: Post-Training-Associated Differences. V2 shifts the comparison from tasks within a model to Base and Instruct variants within the same model family. We compare Base and Instruct variants in

Table 2: Summary of the main findings across the four experimental stages.

Question	Finding	Key Evidence
Do Reader and Self outputs covary?	Yes	Reader–Self change coupling for Clinical–Neutral contrasts was positive across all Primary conditions ($r_\Delta = 0.56$–0.93).
Do Reader and Self rely on the same internal mechanisms?	Partially	Affect-relevant information was decodable from both tasks, but direct cross-decoding and Reader→Self causal transfer provided limited evidence for mechanistic identity.
Is affect-relevant information lost in Instruct models?	No	Affect-relevant information remained decodable in Instruct models, while representation geometry and aspects of causal profiles differed between Base and Instruct models.
Does decodable information causally control Self-Report?	Not necessarily	The State-Induction Gate yielded `NO_GO` for both V/A; subsequent analyses provided only exploratory evidence for generation-stage-dependent causal sensitivity.

affect-relevant representation, Reader–Self sharing, causal utilization, and output readout. Because Base and Instruct models can differ in their sensitivity to prompt format, the primary comparison uses the Matched-Plain condition to minimize prompt-format differences. We interpret the resulting contrasts as *post-training-associated differences*, rather than as causal effects of post-training itself.

V3: Spatiotemporal Causal Localization. V3 focuses on Instruct models and asks where and when affect-relevant information is causally utilized during generation. We examine causal effects across layers and generation stages, using Qwen for discovery analyses and Llama, Gemma, and OLMo for prespecified replication analyses.

V3 uses a prespecified gate before confirmatory causal mapping. We first evaluate whether the candidate intervention satisfies criteria for *Sufficiency*, *Specificity*, and *Endogenous Relevance*, collectively termed the **State Induction Gate**. Subsequent causal mapping analyses are treated as confirmatory only if these criteria are satisfied; otherwise, they are interpreted as exploratory or descriptive. Full definitions, intervention procedures, decision criteria, and stage-specific analyses are provided in Appendix P.

4 Results

Table 2 summarizes the key results across the four stages. We report here only the findings that directly inform the central mechanistic conclusions of the study. Complete model-family-specific results and control analyses are provided in Appendix E, I, M, and Q.

4.1 Behavioral: Reader and Self Strongly Covary

We first examined whether Reader Prediction and Self-Report covary at the behavioral level under controlled stimulus manipulations. For the 192 Clinical–Neutral matched pairs in AIPsy-Affect, we computed

$$\Delta R_i = R(x_i^C) - R(x_i^N), \qquad \Delta S_i = S(x_i^C) - S(x_i^N).$$

Across all Primary model conditions for Valence and Arousal, the Pearson correlations were positive, ranging from

$$r_\Delta = 0.56\text{–}0.93.$$

Thus, Reader–Self behavioral coupling was consistently observed across model families and across Base and Instruct variants: stimuli that induced larger changes in Reader Prediction also tended to induce larger changes in Self-Report.

However, this coupling did not imply human-like affective correspondence. Correspondence with human VAD ratings in EmoBank, Clinical–Neutral Sensitivity, strict Neutral–Moderate–Clinical Dose–Response, and Clinical vs. Complex Neutral Specificity varied substantially across models, tasks, and affective axes. Moreover, only a limited subset of conditions satisfied the strict Dose–Response criterion.

The most robust conclusion supported by the Behavioral Stage is therefore Controlled Stimulus Change $\rightarrow$ Coupled Reader–Self Change, while Reader–Self Coupling $\neq$ Human-Affect Correspondence. Strong output-level covariation alone is insufficient to establish shared internal mechanisms. Full results are provided in Appendix E.

4.2 V1: Shared Information Does Not Imply a Shared Causal Code

V1 examined whether the Behavioral coupling between Reader and Self is supported by identical internal representations or causal mechanisms. Using human Reader Valence/Arousal from EmoBank as external targets, layer-wise probing showed that affect-relevant information could be decoded on held-out data from both Reader and Self hidden states across Base and Instruct variants of all four model families. In many conditions, the peak depths for Reader and Self were also close.

However, decoding the same information does not imply a shared coordinate system. Representational Similarity Analysis showed positive correspondence between Reader and Self geometries, whereas direct cross-decoding was limited in many conditions and improved after Procrustes alignment. Thus, Reader and Self may partially share relational geometry without representing affect-relevant information in directly interchangeable raw coordinate systems.

Semantic controls on AIPsy-Affect further showed a consistent decline in decoding performance under Word Shuffle, making a simple bag-of-words lexical-shortcut account of V1 decodability less plausible. Because the numbers of nonfallback samples available for the Paraphrase and Outcome Reversal controls were small, however, we treat these analyses as supplementary evidence.

We next evaluated causal relationships using hidden-state interventions. Reader and Self exhibited positively correlated layer-wise causal profiles, and their causal peaks occurred at the same or nearby layers in many models. However, cosine similarity between intervention directions was model-dependent. More importantly, cross-task interventions that transferred Reader-derived Clinical–Neutral hidden-state differences into Self did not exhibit robust specificity relative to the random-transfer control after FDR correction.

Thus, Shared / Alignable Information $\not\Rightarrow$ Causally Interchangeable Code.

Task $\times$ SiteType interactions were also observed in some models, but this pattern did not generalize across all model families. We therefore do not conclude that Reader and Self are implemented by two fully independent circuits. Instead, V1 supports an intermediate structure in which affect-relevant information and layer-level causal sensitivity are partially shared, while direct causal interchangeability is not established. Full methodological details and results are provided in Appendix H and Appendix I.

4.3 V2: Post-Training-Associated Differences Are Not Explained by Simple Erasure

V2 shifted the primary comparison from Reader–Self to Base–Instruct. Because Base and Instruct models are typically evaluated using different prompt formats, the Primary comparison used the Matched-Plain condition, in which both models received plain-text prompts.

We first compared representational geometry. The mean Base–Instruct Procrustes distortion was

$$1.455 \qquad (95\%\,\mathrm{CI} = [0.639, 2.813])$$

for Reader and

$$2.872 \qquad (95\%\,\mathrm{CI} = [0.660, 6.995])$$

for Self, indicating representational-geometric disparities for both tasks. We treat these quantities as descriptive cross-family summaries and do not interpret the exclusion of zero from the distance confidence intervals as a null-hypothesis test.

Importantly, V1 showed that affect-relevant information remained decodable from Instruct models. Thus, the Base–Instruct difference is more consistent with Information Remains Decodable + Its Organization Differs than with a simple account in which Information Is Erased.

However, this difference did not correspond to a simple "deeper relocation." For Valence Reader, the decodability peak instead shifted toward earlier layers,

$$\Delta d_D^* = -0.119, \qquad 95\%\,\mathrm{CI} = [-0.219, -0.033],$$

while no consistent shift was observed for the other task / axis combinations. Changes in Reader–Self sharing also had cross-family confidence intervals that included zero for both Valence and Arousal, providing no support for uniform task separation.

For the causal profiles, only the Valence Post-training $\times$ Task interaction remained supported after the Primary FDR correction:

$$\beta = 1.63 \times 10^{-4}, \qquad q = 0.044.$$

By contrast, neither the Post-training $\times$ Depth interaction nor the three-way interaction was supported. Evidence for causal reorganization was therefore limited to a task-dependent Valence component rather than a family-general relocation of causal leverage across depth.

Finally, we patched Base-derived representations into Instruct models and evaluated recovery toward the Base output distribution. Under Matched-Plain prompting, Recovery AUC for Qwen was 0.189 for Reader and 0.181 for Self. In contrast, Llama was close to zero, Gemma was negative for Reader and approximately zero for Self, and OLMo was negative for both Reader and Self, indicating strongly family-dependent recoverability. The Self–Reader recovery asymmetry was also not supported:

$$\Delta\mathrm{AUC} = 0.011, \qquad 95\%\,\mathrm{CI} = [-0.010, 0.038].$$

Taken together, V2 does not support reducing the Base–Instruct difference to any single account based on Erasure, Uniform Relocation, Uniform Reader–Self Separation, or a Single Recoverable Bottleneck. More narrowly, the results are consistent with distributed post-training-associated reorganization spanning representational geometry, task-dependent causal sensitivity, and model-specific recoverability. Because Base and Instruct models are separately obtained artifacts rather than conditions generated by a randomized post-training intervention, these differences should not be interpreted as causal effects of post-training itself. Full methodological details and results are provided in Appendix L and Appendix M.

4.4 V3: Decodability Does Not Guarantee Robust Causal Control

V3 restricted the analysis to Instruct models and examined whether affect-relevant information exerts measurable causal leverage over Self-Report generation. Using Qwen 2.5 Instruct as the Discovery model, we first applied a State Induction Gate to Reader-grounded affect directions. The Gate comprised Sufficiency, Specificity, Endogenous Relevance, and Topic Control criteria.

The key result was that the overall Gate yielded `NO_GO` for both Valence and Arousal. Although the Topic Control passed, Sufficiency, Specificity, and Endogenous Relevance did not satisfy their prespecified criteria. Thus,

$$\text{Decodable} \not\Rightarrow \text{Robustly Steerable},$$

and the Reader-grounded affect directions identified here cannot be regarded as robust and specific causal controllers of Self-Report.

Analyses conducted after the Gate failure are interpreted as exploratory and descriptive evidence. In the subsequent analyses, the teacher-forced layer $\times$ generation-stage map for Qwen showed that the decodability peak occurred at a relative depth of $d^*_D \simeq 0.85$, whereas the causal-displacement peak occurred earlier, at $d^*_C \simeq 0.74$–0.78. After freezing the direction of this spatial dissociation following Qwen Discovery, Llama 3.2 and OLMo 2 satisfied the directional criterion for Valence, whereas the pattern was not consistent for Gemma 3 or for Arousal.

Moreover, neither Sufficiency at the fixed intervention sites nor mediated attenuation induced by affect-related subspace removal satisfied the cross-family criterion across the three replication families. Thus, the results did not support family-general Sufficiency or the stronger claim that the identified subspace serves as a major endogenous mediator of natural Self-Report shifts.

After Qwen Discovery, we froze the direction of this spatial dissociation before examining outcomes in the replication families. For Valence, Llama 3.2 and OLMo 2 satisfied the frozen directional criterion, whereas the pattern was not consistent for Gemma 3 or for Arousal.

Sufficiency at the frozen site did not satisfy the replication criterion in any of the three families. Likewise, mediated attenuation induced by removal of the affect-related subspace did not satisfy the

cross-family replication criterion. Thus, the subspace identified here was not established as a major mediator of naturally occurring Self-Report shifts.

By contrast, the temporal contrast between the target generation stage frozen from Qwen Discovery and candidate start satisfied the criterion for both Valence and Arousal in all three replication families—Llama, Gemma, and OLMo. This provides an exploratory replication signal that local interventional sensitivity need not be uniform throughout the generation sequence and may vary across generation stages.

However, this temporal result does not establish a single family-general causal mechanism. First, the State Induction Gate yielded `NO_GO`. Second, the H1–H3 replication criteria were not consistently satisfied. Third, the generation-stage analysis was conducted along a teacher-forced trajectory defined by a fixed canonical candidate rather than along the model's free-generation trajectory.

V3 therefore supports the more limited conclusion that decodability, local sensitivity to external intervention, and endogenous relevance during natural computation must be distinguished. Full methodological details and results are provided in Appendix P and Appendix Q.

5 Discussion

Our results reveal a progressive dissociation between behavioral correspondence, representational sharing, and causal utilization. Reader Prediction and Self-Report were strongly coupled behaviorally, and affect-relevant information was decodable from both tasks. However, representational commonality was only partial, and cross-task causal interchangeability was not robustly established. Thus, behavioral coupling and the presence of decodable affect-relevant information do not by themselves imply that the two tasks rely on the same internal mechanism.

The Base–Instruct and generation-stage analyses further reinforce the distinction between the presence of affect-relevant information and how that information is organized and utilized. Affect-relevant information remained decodable in Instruct models despite differences in representational organization and aspects of causal utilization, arguing against a simple account in which Base–Instruct differences reflect the disappearance of affect-relevant information. Similarly, although affect-relevant information remained decodable during Self-Report generation, our interventions did not provide robust evidence that it exerts stable and specific endogenous causal control over Self-Report. Exploratory analyses suggested that interventional sensitivity may vary across layers and generation stages, but did not establish a single family-general causal mechanism. Because the Base–Instruct comparison is observational, these differences should be interpreted as post-training-associated rather than as causal effects of post-training itself.

Taken together, our findings support an intermediate account of the relationship between Reader Prediction and Self-Report: **the two tasks partially share affect-relevant information, while the way this information is represented and causally utilized need not be identical.** This account is inconsistent with both a simple fully shared mechanism, in which a single affective representation is read out identically by both tasks, and a fully independent account. Instead, our results indicate that decodability, representational sharing, and causal utilization are distinct properties that need not coincide. Mechanistic claims about affect-related behavior should therefore not be inferred from behavioral correspondence or decodability alone.

5.1 Implications for Psychological Guardrails

This distinction has implications for LLMs used in affect-sensitive settings such as psychological support. Such systems may need to preserve capabilities relevant to predicting human affective responses while limiting self-referential affective expressions that imply that the model itself has emotions or that may intensify users' emotional attachment to the system. Uniformly suppressing affect-relevant internal information may therefore be undesirable if some of that information supports both Reader Prediction and Self-Report.

At the same time, our results do not support the stronger claim that the two tasks utilize affect-relevant information through identical mechanisms. This leaves open the possibility of more selective interventions that preserve information relevant to predicting human affective responses while regulating how that information is utilized for affective Self-Report. Our study does not establish

such a guardrail, but suggests that its design may need to target **how affect-relevant information is utilized**, rather than simply **whether that information is present**.

6 LIMITATIONS

This study has several limitations concerning the scope and interpretation of our findings. Self-Report is operationalized as an elicited numerical model output and should not be interpreted as evidence of subjective emotional experience; likewise, Reader Prediction should not be equated with human cognitive empathy. Our experiments are also limited to four open-weight model families in the 1–1.5B parameter range, so the findings may not generalize to larger or closed-weight models. In addition, the Base–Instruct comparison is observational rather than a randomized post-training intervention, and should therefore be interpreted as identifying post-training-associated differences rather than causal effects of post-training itself.

Several methodological limitations further constrain the evidence. Behavioral/V1 and V2/V3 use different candidate spaces, so absolute affect scores should not be compared across these stages. V3 analyzes internal computation along teacher-forced trajectories with a fixed canonical response rather than unconstrained free generation, leaving open whether the same spatiotemporal patterns arise during natural generation. Finally, some analyses involve limited sample sizes and substantial cross-family heterogeneity. Accordingly, the absence of a statistically robust effect should not be interpreted as evidence that affect-relevant information is absent or that Reader Prediction and Self-Report rely on completely independent mechanisms.

7 CONCLUSION

Reader Prediction and Self-Report showed strong behavioral coupling, and affect-relevant information was decodable from both tasks, but evidence for shared representations and cross-task causal utilization was more limited. Base–Instruct and generation-stage analyses further showed that decodability does not imply invariant representational organization or robust endogenous causal control. Together, these findings support an intermediate account in which Reader Prediction and Self-Report share some affect-relevant information without necessarily representing or utilizing that information in the same way. This distinction motivates selective interventions that preserve capabilities relevant to predicting human affective responses while regulating how affect-relevant information is utilized for self-referential affective expression.

REPRODUCIBILITY STATEMENT

The Appendix documents the datasets, model revisions, prompts, candidate spaces, Sequence-Likelihood measurements, hidden-state locations, random seeds, statistical procedures, and Primary/Secondary analysis definitions. The Matched-Plain condition is the Primary Base–Instruct comparison, and V3 separates Discovery from cross-family replication, with analyses after the State Induction Gate yielded `NO_GO` treated as exploratory/descriptive.

AI USE STATEMENT

We used generative AI tools for manuscript organization and language editing, organizing related-work materials, reviewing and debugging experimental code, and supplementary brainstorming on research questions and experimental design.

We did not use generative AI to generate or modify experimental data, determine experimental or statistical results, or adopt research claims without author verification. All AI-assisted text, claims, code, and related-work information were checked against the original literature, experimental results, and implementations and revised where necessary; experimental code was additionally verified against execution results, tests, and output files. The authors take full responsibility for all final content.

REFERENCES

John W Ayers, Adam Poliak, Mark Dredze, Eric C Leas, Zechariah Zhu, Jessica B Kelley, Dennis J Faix, Aaron M Goodman, Christopher A Longhurst, Michael Hogarth, et al. Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. *JAMA internal medicine*, 183(6):589–596, 2023.

Amin Banayeeanzade, Ala N Tak, Fatemeh Bahrani, Anahita Bolourani, Leonardo Blas, Emilio Ferrara, Jonathan Gratch, and Sai Praneeth Karimireddy. Psychological steering in llms: An evaluation of effectiveness and trustworthiness. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1719–1771, 2026.

Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, 57(1):289–300, 1995.

Felix Jedidja Binder, James Chua, Tomek Korbak, Henry Sleight, John Hughes, Robert Long, Ethan Perez, Miles Turpin, and Owain Evans. Looking inward: Language models can learn about themselves by introspection. In *International Conference on Learning Representations*, volume 2025, pp. 3710–3756, 2025.

Sven Buechel and Udo Hahn. Emobank: Studying the impact of annotation perspective and representation format on dimensional emotion analysis. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pp. 578–585, 2017.

Collin Burns, Haotian Ye, Dan Klein, and Jacob Steinhardt. Discovering latent knowledge in language models without supervision. *arXiv preprint arXiv:2212.03827*, 2022.

Mark H Davis. Measuring individual differences in empathy: Evidence for a multidimensional approach. *Journal of personality and social psychology*, 44(1):113, 1983.

Jean Decety and Philip L Jackson. The functional architecture of human empathy. *Behavioral and cognitive neuroscience reviews*, 3(2):71–100, 2004.

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

Zohar Elyoseph, Dorit Hadar-Shoval, Kfir Asraf, and Maya Lvovsky. Chatgpt outperforms humans in emotional awareness evaluations. *Frontiers in psychology*, 14:1199058, 2023.

Cathy Mengying Fang, Auren R Liu, Valdemar Danry, Eunhae Lee, Samantha WT Chan, Pat Pataranutaporn, Pattie Maes, Jason Phang, Michael Lampe, Lama Ahmad, et al. How ai and human behaviors shape psychosocial effects of extended chatbot use: A longitudinal randomized controlled study. *arXiv preprint arXiv:2503.17473*, 2025.

Javier Ferrando, Oscar Obeso, Senthooran Rajamanoharan, and Neel Nanda. Do i know this entity? knowledge awareness and hallucinations in language models. In *International Conference on Learning Representations*, volume 2025, pp. 77965–78000, 2025.

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

Harry Haroutioun Haladjian and Carlos Montemayor. Artificial consciousness and the consciousness-attention dissociation. *Consciousness and cognition*, 45:210–225, 2016.

John Hewitt and Percy Liang. Designing and interpreting probes with control tasks. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (emnlp-ijcnlp)*, pp. 2733–2743, 2019.

Toshiyuki Himichi, Hidekazu Osanai, Takayuki Goto, Hiroyo Fujita, Yuta Kawamura, Mark H Davis, and Michio Nomura. Development of a japanese version of the interpersonal reactivity index. *Shinrigaku Kenkyu: The Japanese Journal of Psychology*, 88(1):61–71, 2017.

Jen-tse Huang, Man Ho Lam, Eric John Li, Shujie Ren, Wenxuan Wang, Wenxiang Jiao, Zhaopeng Tu, and Michael R Lyu. Emotionally numb or empathetic? evaluating how llms feel using emotionbench. *arXiv preprint arXiv:2308.03656*, 2023.

Michael Keeman. Aipsy-affect: A keyword-free clinical stimulus battery for mechanistic interpretability of emotion in language models. *arXiv preprint arXiv:2604.23719*, 2026.

Nikolaus Kriegeskorte, Marieke Mur, and Peter A Bandettini. Representational similarity analysis-connecting the branches of systems neuroscience. *Frontiers in systems neuroscience*, 2:249, 2008.

Keido Labs. aipsy-affect (revision e438b14), 2026. URL `https://huggingface.co/datasets/keidolabs/aipsy-affect`.

Linnea Laestadius, Andrea Bishop, Michael Gonzalez, Diana Illenčík, and Celeste Campos-Castillo. Too human and not human enough: A grounded theory analysis of mental health harms from emotional dependence on the social chatbot replika. *new media & society*, 26(10):5923–5941, 2024.

Han Li, Renwen Zhang, Yi-Chieh Lee, Robert E Kraut, and David C Mohr. Systematic review and meta-analysis of ai-based conversational agents for promoting mental health and well-being. *NPJ Digital Medicine*, 6(1):236, 2023a.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Inference-time intervention: Eliciting truthful answers from a language model. *Advances in neural information processing systems*, 36:41451–41530, 2023b.

Jack Lindsey. Emergent introspective awareness in large language models. *arXiv preprint arXiv:2601.01828*, 2026.

Xiaochen Luo, Zixuan Wang, Jacqueline L Tilley, Sanjeev Balarajan, Ukeme-Abasi Bassey, and Choi Ieng Cheang. Seeking emotional and mental health support from generative ai: mixed-methods study of chatgpt user experiences. *JMIR Mental Health*, 12(1):e77951, 2025.

Nicolas Martorell and Bruno Bianchi. Quantitative introspection in language models: Tracking emotive states across conversation. *arXiv preprint arXiv:2603.18893*, 2026.

Albert Mehrabian and James A Russell. The basic emotional impact of environments. *Perceptual and motor skills*, 38(1):283–301, 1974.

Carlos Montemayor, Jodi Halpern, and Abrol Fairweather. In principle obstacles for empathic ai: why we can't replace human empathy in healthcare. *AI & society*, 37(4):1353–1359, 2022.

Team OLMo, Pete Walsh, Luca Soldaini, Dirk Groeneveld, Kyle Lo, Shane Arora, Akshita Bhagia, Yuling Gu, Shengyi Huang, Matt Jordan, Nathan Lambert, Dustin Schwenk, Oyvind Tafjord, Taira Anderson, David Atkinson, Faeze Brahman, Christopher Clark, Pradeep Dasigi, Nouha Dziri, Michal Guerquin, Hamish Ivison, Pang Wei Koh, Jiacheng Liu, Saumya Malik, William Merrill, Lester James V. Miranda, Jacob Morrison, Tyler Murray, Crystal Nam, Valentina Pyatkin, Aman Rangapur, Michael Schmitz, Sam Skjonsberg, David Wadden, Christopher Wilhelm, Michael Wilson, Luke Zettlemoyer, Ali Farhadi, Noah A. Smith, and Hannaneh Hajishirzi. 2 OLMo 2 Furious, 2024a. URL `https://arxiv.org/abs/2501.00656`.

Team OLMo, Pete Walsh, Luca Soldaini, Dirk Groeneveld, Kyle Lo, Shane Arora, Akshita Bhagia, Yuling Gu, Shengyi Huang, Matt Jordan, et al. 2 olmo 2 furious. *arXiv preprint arXiv:2501.00656*, 2024b.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li,

Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report, 2025. URL https://arxiv.org/abs/2412.15115.

Benjamin Reichman, Adar Avsian, and Larry Heck. Emotions where art thou: Understanding and characterizing the emotional latent space of large language models. In *International Conference on Learning Representations*, volume 2026, pp. 49924–49947, 2026.

Yossi Rubner, Carlo Tomasi, and Leonidas J Guibas. The earth mover's distance as a metric for image retrieval. *International journal of computer vision*, 40(2):99–121, 2000.

Peter H Schönemann. A generalized solution of the orthogonal procrustes problem. *Psychometrika*, 31(1):1–10, 1966.

Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural machine translation of rare words with subword units. In *Proceedings of the 54th annual meeting of the association for computational linguistics (volume 1: long papers)*, pp. 1715–1725, 2016.

Simone G Shamay-Tsoory. The neural bases for empathy. *The Neuroscientist*, 17(1):18–24, 2011.

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

Tania Singer and Claus Lamm. The social neuroscience of empathy. *Annals of the new York Academy of Sciences*, 1156(1):81–96, 2009.

Shashwat Singh, Tal Linzen, and Shauli Ravfogel. Can llms introspect? a reality check. *arXiv preprint arXiv:2605.26242*, 2026.

Ala N Tak, Amin Banayeeanzade, Anahita Bolourani, Mina Kian, Robin Jia, and Jonathan Gratch. Mechanistic interpretability of emotion inference in large language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 13090–13120, 2025.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, et al. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*, 2025.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel Bowman. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. *Advances in Neural Information Processing Systems*, 36:74952–74965, 2023.

Amy Beth Warriner, Victor Kuperman, and Marc Brysbaert. Norms of valence, arousal, and dominance for 13,915 english lemmas. *Behavior research methods*, 45(4):1191–1207, 2013.

Laura Weidinger, Jonathan Uesato, Maribeth Rauh, Conor Griffin, Po-Sen Huang, John Mellor, Amelia Glaese, Myra Cheng, Borja Balle, Atoosa Kasirzadeh, et al. Taxonomy of risks posed by language models. In *Proceedings of the 2022 ACM conference on fairness, accountability, and transparency*, pp. 214–229, 2022.

Xiuwen Wu, Hao Wang, Zhiang Yan, Xiaohan Tang, Pengfei Xu, Wai-Ting Siok, Ping Li, Jia-Hong Gao, Bingjiang Lyu, and Lang Qin. Decoding and controlling emotion in llms through human-aligned representational geometry with enhanced interpretability. *Computers in Human Behavior*, pp. 109051, 2026.

Tengfei Yu, Siyu Pan, Caoyun Fan, Siyang Luo, Yaohui Jin, and Binglei Zhao. Can large language models exhibit cognitive and affective empathy as humans? *Computers in Human Behavior: Artificial Humans*, pp. 100233, 2025.

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

Shaolei Zhang, Tian Yu, and Yang Feng. Truthx: Alleviating hallucinations by editing large language models in truthful space. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8908–8949, 2024.

Shuai Zhang, Yuxing Qian, Zhizhen Yao, Zhenni Ni, and Yang Zhang. From approach to avoidance: How ai agent cognitive and affective empathy elicits the uncanny valley effect. *Telematics and Informatics*, pp. 102313, 2025.

Appendix

A Background

Large language model (LLM)-based conversational AI systems are increasingly used not only to assist with tasks such as text generation, information retrieval, summarization, and question answering, but also in contexts involving users' psychological well-being and interpersonal lives, including seeking advice, developing self-understanding, alleviating loneliness, and supporting decision-making. AI-based conversational agents have been widely explored as a means of supporting mental health and well-being (Li et al., 2023a). Recent work has further shown that users seeking emotional and psychological support from generative AI engage with LLMs for a broad range of purposes beyond psychoeducation and therapy-related concerns, including companionship, interpersonal relationships, well-being, and decision-making (Luo et al., 2025).

While these uses offer new forms of support, they also introduce risks specific to human–AI interaction. The risks associated with language models extend beyond content-level harms, such as misinformation and discriminatory outputs, to harms that may emerge through sustained interactions between humans and AI systems (Weidinger et al., 2022). Of particular concern is emotional dependence, in which users perceive AI systems as social entities with emotions and desires and become psychologically invested in maintaining their relationships with them (Laestadius et al., 2024). Research on long-term chatbot use has also examined how factors such as usage frequency and trust in AI are associated with psychosocial outcomes and problematic use (Fang et al., 2025).

Therefore, ensuring the safety of AI systems designed to provide psychological support should not simply amount to reducing users' interactions with AI. Rather, such systems should preserve the benefits of psychological support while preventing users from becoming excessively involved in their relationships with AI and enabling them to maintain their autonomy.

In this work, we refer to mechanisms that regulate AI responses toward this goal as psychological guardrails. Psychological guardrails are not limited to mechanisms that simply refuse harmful content. Instead, they aim to selectively regulate responses that may foster excessive dependence or emotional involvement, while preserving the functions that make AI systems useful as a source of psychological support.

One particularly important factor in designing such guardrails is empathetic responding.

In psychology, empathy has been conceptualized not as a unitary capacity, but as a multidimensional construct comprising multiple components (Davis, 1983; Decety & Jackson, 2004; Himichi et al., 2017). A common distinction is between cognitive empathy, which involves understanding or inferring another person's perspective, intentions, emotions, and mental states, and affective empathy, which involves experiencing an emotional response to another person's emotional state (Singer & Lamm, 2009; Shamay-Tsoory, 2011).

This distinction is also important when considering AI systems for psychological support. The ability to accurately understand another person's emotions and circumstances and to incorporate that understanding into a response can be valuable for providing psychological support. In contrast, responses that appear to excessively mirror a user's emotions, or that imply that the AI itself experiences similar emotions, may shape how users perceive their relationship with the AI and potentially contribute to excessive emotional involvement (Laestadius et al., 2024; Zhang et al., 2025).

Designing effective psychological guardrails therefore requires understanding the extent to which two processes—inferring another person's emotional state and reporting the model's own affective response to the same stimulus—rely on shared information, representations, and causal pathways within an LLM. If these processes are supported by distinct internal mechanisms, it may be possible

to selectively intervene on computations associated with excessive self-directed affective responses while preserving the capabilities required to understand others. Conversely, if they rely heavily on shared representations or causal pathways, interventions targeting one process may inadvertently impair the model's capacity to provide psychological support.

This question cannot be resolved by examining generated responses alone. A growing body of behavioral evaluations has demonstrated that LLMs can produce responses that appear empathetic (Ayers et al., 2023; Elyoseph et al., 2023; Huang et al., 2023; Yu et al., 2025). However, observing an empathetic response does not reveal whether the underlying computation involves inferring another person's state, nor how that computation relates to affective responses that the model generates about itself.

Moreover, it is crucial to distinguish between information being represented within an LLM and that information being causally used to produce its final output.

Even if valence and arousal can be decoded from a hidden state with high accuracy,

$$\text{Information is decodable} \not\Rightarrow \text{Information is causally used.}$$

Similarly, even when a model reports an affective state about itself, such a report does not necessarily constitute a direct or faithful readout of its internal representations. Prior work has shown that, in LLMs, information represented internally may not align with the model's surface-level reports (Burns et al., 2022; Turpin et al., 2023). Moreover, post-training may alter the mapping between existing internal representations and the model's output policy while leaving those representations largely intact (Sharma et al., 2024; Ferrando et al., 2025).

We therefore operationally separate two computations elicited by the same external affective stimulus.

First, in **Reader Prediction**, the model is asked to predict how a human reader would feel after reading a given text:

$$x \rightarrow \text{prediction of a human reader's affect.}$$

Second, in **Self-Report**, the model is presented with the same text and asked to numerically report an affective state about itself:

$$x \rightarrow \text{model-generated affective report about itself.}$$

Reader Prediction operationalizes the computation involved in inferring another individual's affective state, whereas Self-Report operationalizes the model-generated affective response about itself to the same stimulus.

This Reader–Self distinction is inspired by the psychological distinction between cognitive and affective empathy. However,

$$\text{Reader Prediction} \neq \text{human cognitive empathy},$$

and

$$\text{Self-Report} \neq \text{human affective empathy}.$$

Throughout this work, we therefore treat Reader Prediction and Self-Report not as direct instantiations of human empathic processes, but as computational operationalizations for studying LLMs.

In particular, Self-Report does not presuppose that the model possesses human-like emotional experiences, subjective states, consciousness, or sentience. Accordingly, our focus is not the ontological question of whether AI "has" emotions. Rather, we study a mechanistic question: how affect-relevant information is represented within LLMs, how it is used for Reader Prediction and Self-Report, and how the relationship between these processes changes through post-training and the generation process.

The goal of addressing this question is not to determine whether AI possesses subjective feelings. Instead, we aim to provide a mechanistic foundation for designing psychological guardrails that can more selectively regulate responses that may promote excessive emotional alignment or dependence, while preserving the model's ability to understand others' emotions and provide psychological support.

We frame this problem as the following computational sequence:

Stimulus → Affect-relevant Representation → Task-dependent Computation → Causal Utilization → Report

and investigate it through four stages: **Behavioral**, **V1**, **V2**, and **V3**.

B Related Work

B.1 Conversational AI for Psychological Support and Human–AI Dependence

Conversational AI has been extensively studied as a technology for supporting mental health and well-being. Systematic reviews and meta-analyses of AI-based conversational agents have characterized both their potential and limitations in promoting psychological health and well-being (Li et al., 2023a). With the widespread adoption of generative AI, users have also increasingly turned to systems such as ChatGPT for emotional and mental-health support. Such use extends beyond psychoeducation and therapy-related concerns to a broader range of purposes, including companionship, interpersonal relationships, well-being, and decision-making (Luo et al., 2025).

At the same time, sustained relationship formation with AI systems may introduce distinct risks. Laestadius et al. (2024) report emotional dependence among Replika users, whereby users attribute emotions and desires to the AI and become psychologically invested in maintaining their relationship with it (Laestadius et al., 2024). Fang et al. (2025) further examine extended chatbot use in a longitudinal study, investigating the relationships between usage patterns, trust in AI, and psychosocial outcomes (Fang et al., 2025).

These concerns also appear in taxonomies of risks associated with language models, which extend beyond content-level harms such as misinformation to risks arising from human–AI interaction itself (Weidinger et al., 2022).

Taken together, these studies motivate AI-side mechanisms that mitigate excessive dependence and emotional involvement while preserving the capabilities that make conversational AI useful for psychological support, rather than simply removing such supportive capabilities altogether. However, we still lack a mechanistic understanding of where and how such control should be applied within the model's internal computations.

B.2 Cognitive and Affective Empathy in Psychology

In psychology, empathy has been conceptualized not as a unitary capacity, but as a multidimensional construct comprising multiple components.

Through the Interpersonal Reactivity Index (IRI), Davis (1983) proposed a multidimensional model of empathy encompassing components such as perspective taking, empathic concern, and personal distress. Decety & Jackson (2004) similarly conceptualized human empathy as involving multiple interacting processes, including affective sharing, self–other distinction, and perspective taking.

Neurocognitive research has further distinguished, at least partially, between systems involved in affective responses to others' emotional states and cognitive processes involved in understanding others' mental states and perspectives (Singer & Lamm, 2009; Shamay-Tsoory, 2011). This multidimensional view has also been supported in Japanese populations through the development and validation of a Japanese version of the IRI, which provides a framework for measuring both cognitive and affective dimensions of empathy (Himichi et al., 2017).

The distinction between Reader Prediction and Self-Report in our study is inspired by this distinction between the cognitive and affective dimensions of empathy. Importantly, however, we do not equate Reader Prediction with human cognitive empathy, nor Self-Report with human affective empathy.

B.3 Empathic AI and Behavioral Evaluations of Empathy in LLMs

Whether artificial systems can possess states equivalent to human empathy has been the subject of conceptual and philosophical debate, distinct from behavioral evaluations of their observable capabilities.

Haladjian & Montemayor (2016) discuss the relationships among attention, consciousness, and emotion in artificial systems, arguing for a distinction between the simulation of emotion or empathy by AI and human phenomenal experience. Montemayor et al. (2022) similarly discuss empathic AI in healthcare, highlighting fundamental difficulties in equating an AI system's ability to infer a patient's state and generate an appropriate response with empathy as experienced by humans.

These works do not empirically establish the absence of subjective feelings in AI systems; rather, they provide theoretical arguments concerning the extent to which concepts such as empathy can be meaningfully applied to artificial systems. Our work does not aim to resolve this ontological debate. Instead, we restrict our analysis to observable representations and causal computations within LLMs.

In parallel, a growing body of empirical work has investigated whether LLMs exhibit behaviors corresponding to cognitive and affective empathy. Yu et al. (2025) evaluate the behavioral structure of cognitive and affective empathy in GPT-4 and Llama-3 using the Interpersonal Reactivity Index and the Basic Empathy Scale. Zhang et al. (2025) report that cognitive and affective empathy exhibited by AI agents can differentially influence users' approach–avoidance responses.

Beyond these studies, Ayers et al. (2023) compare responses from ChatGPT and physicians to patient questions and report that ChatGPT's responses were rated favorably in terms of both quality and empathy. Elyoseph et al. (2023) evaluate ChatGPT using the Levels of Emotional Awareness Scale and report high emotional-awareness performance. EmotionBench evaluates stimulus-induced affective responses of LLMs across a broad range of emotion-eliciting situations, analyzing both agreements and discrepancies between model and human responses (Huang et al., 2023).

Collectively, these studies demonstrate that LLMs can generate sophisticated behaviors that appear empathetic or affective. However,

$$\text{Behavioral similarity} \neq \text{Mechanistic identity},$$

and observing a model's final generated response alone is insufficient to identify the internal representations or causal mechanisms underlying that behavior.

B.4 Affect-Related Representations in LLMs

Hidden-state probing studies have shown that information related to sentiment and emotion can be linearly decoded from the internal representations of LLMs.

Di Palma et al. analyze hidden states across the LLaMA family and show that the decodability of sentiment- and emotion-related information varies across layers (Di Palma et al., 2025). Tak et al. investigate internal representations associated with emotion inference and demonstrate that causal interventions on internal concepts can alter emotion-related outputs (Tak et al., 2025). Zhang and Zhong further show that emotion-related information exhibits structure not only across layers but also along the token sequence (Zhang & Zhong, 2025).

Recent work further suggests that emotion-related representations may exhibit low-dimensional latent structure and geometries corresponding to dimensions such as valence and arousal. Reichman et al. investigate emotional latent spaces within LLMs (Reichman et al., 2026), while Wu et al. (2026) identify human-aligned geometries of emotion representations and explore the use of the resulting directions for controlling emotion-related behavior.

Banayeeanzade et al. (2026) compare psychological steering through prompting, fine-tuning, and representation engineering for emotion and personality, evaluating both the controllability afforded by activation-space interventions and their side effects on other model behaviors.

Taken together, these studies provide empirical support for the relationship

$$\text{emotion-related information} \rightarrow \text{decodable and structured internal representations}.$$

B.5 Decodability, Causal Intervention, and Endogenous Relevance

The fact that information can be decoded from a hidden state does not imply that the model actually uses that information at that location to produce downstream outputs (Hewitt & Liang, 2019).

Similarly, demonstrating that injecting a representation direction can alter model outputs may provide evidence of **sufficiency** (Li et al., 2023b; Tak et al., 2025; Banayeeanzade et al., 2026). How-

ever, this is distinct from **endogenous relevance**: whether the representation is naturally recruited and causally used during the model's processing of an unperturbed stimulus.

Accordingly, we distinguish three levels of evidence:

$$\text{Decodability} \rightarrow \text{Sufficiency} \rightarrow \text{Endogenous Relevance}.$$

Here, decodability asks whether task-relevant information can be recovered from internal states; sufficiency asks whether externally manipulating the corresponding representation can drive the target behavior; and endogenous relevance asks whether the model's naturally occurring computation actually relies on that representation when processing the stimulus.

B.6 Dissociation Between Internal Information and Surface Outputs

Beyond the emotion domain, prior work has demonstrated that information represented internally within an LLM need not align with the reports it ultimately generates.

Burns et al. (2022) show that latent knowledge can potentially be recovered from internal activations independently of a model's explicit answers, motivating a distinction between what the model knows" and what the model says". Turpin et al. (2023) further demonstrate that Chain-of-Thought explanations do not necessarily faithfully reflect the factors that actually drive a model's predictions.

Related work has also investigated whether directly intervening on truthfulness-related internal representations can alter model outputs. Inference-Time Intervention modifies model behavior by intervening on activation directions associated with truthfulness (Li et al., 2023b), while TruthX explores activation editing based on a representation space associated with truthful and deceptive behavior (Zhang et al., 2024).

Although these studies are not specific to emotion, they support the distinction

$$\text{Internal Information} \neq \text{Surface Report},$$

and motivate examining not only whether a representation is present, but also whether and how it is causally involved in producing model outputs.

B.7 Self-Report and LLM Introspection

The relationship between reports that a model generates about itself and its internal states has recently become a direct subject of research on LLM introspection.

Binder et al. (2025) investigate whether models may have privileged access to self-specific information through tasks that require them to predict their own future behavior. Lindsey (2026) examines whether models can report internal manipulations in which representations corresponding to known concepts are directly injected into their activations. Martorell & Bianchi (2026) analyzes the relationship between probe-defined internal states and numerical Self-Reports, investigating whether distributional or logit-based measurements can reveal correspondence with internal states even when greedily decoded reports collapse to a small number of discrete values.

At the same time, Singh et al. (2026) argue that interpreting such phenomena as genuine introspection requires more carefully distinguishing privileged access that cannot be explained by surface-level information from second-order computations about internal representations.

Our work does not aim to resolve this debate by establishing that LLMs engage in genuine introspection, nor do we infer from Self-Report that LLMs experience subjective emotions. Instead, we treat Self-Report strictly as an operational variable: a *model-generated affective report about itself*.

B.8 Reorganization of the Representation–Report Relationship through Post-Training

Behavioral changes induced by post-training need not arise solely from the acquisition or loss of internal information; they may also result from changes in how pre-existing representations are utilized or mapped onto the model's output policy.

Sharma et al. (2024) analyze sycophancy in assistants trained with human feedback and show that model behavior can shift toward responses aligned with human preferences. Ferrando et al. (2025)

demonstrate that entity-recognition-related representation directions identified in a base model also exert causal effects on refusal behavior in its chat-tuned counterpart, suggesting that chat fine-tuning can repurpose pre-existing mechanisms toward a different output policy.

Accordingly, differences between Base and Instruct models should not be reduced to a simple dichotomy of

representation exists vs. representation disappears.

Even when affect-relevant information remains represented after post-training, post-training may alter

- the geometry of its internal representations,
- the degree to which Reader and Self computations share representations,
- the layers at which those representations exert causal influence, and
- the mapping from internal representations to surface-level reports.

Thus, behavioral differences between Base and Instruct models may reflect a reorganization of how affect-relevant representations are computationally utilized, rather than simply their presence or absence.

B.9 Open Questions from Prior Work

Taken together, prior work has established that

1. LLMs can generate behaviors that appear empathetic or affective,
2. emotion-related information can be decoded from hidden states,
3. affective representations may exhibit structured geometry,
4. interventions on internal activations can alter emotion-related outputs,
5. internal representations need not align with surface-level reports, and
6. correspondence between internal states and Self-Reports can sometimes be observed.

However, a key question for designing psychological guardrails remains unresolved: to what extent do the processes of understanding and inferring another person's affective state and reporting an affective response about the model itself to the same stimulus rely on shared internal information, representational geometry, and causal pathways? It also remains unclear how the relationship between these processes is reorganized by post-training-associated changes from Base to Instruct models, and at which layers and stages of the generation process causal leverage over Self-Report emerges.

To address these gaps, we conduct a four-stage analysis.

Our goal is not to determine whether AI systems possess subjective emotional experiences. Rather, we aim to characterize the internal mechanisms linking Reader Prediction and Self-Report, with the longer-term goal of providing a mechanistic basis for more selective psychological guardrails: interventions that preserve the ability to understand others' emotions and provide psychological support while reducing responses that may promote excessive emotional alignment or dependence.

C Overview of the Experimental Framework

The outline of overall framework is shown in Fig. 1.

C.1 Overall Objective and Study Design

The central Research Question (RQ) of this study is:

> *How are affect-relevant internal representations in LLMs related to Self-Report, and how does this relationship vary across tasks, post-training, and stages of computation?*

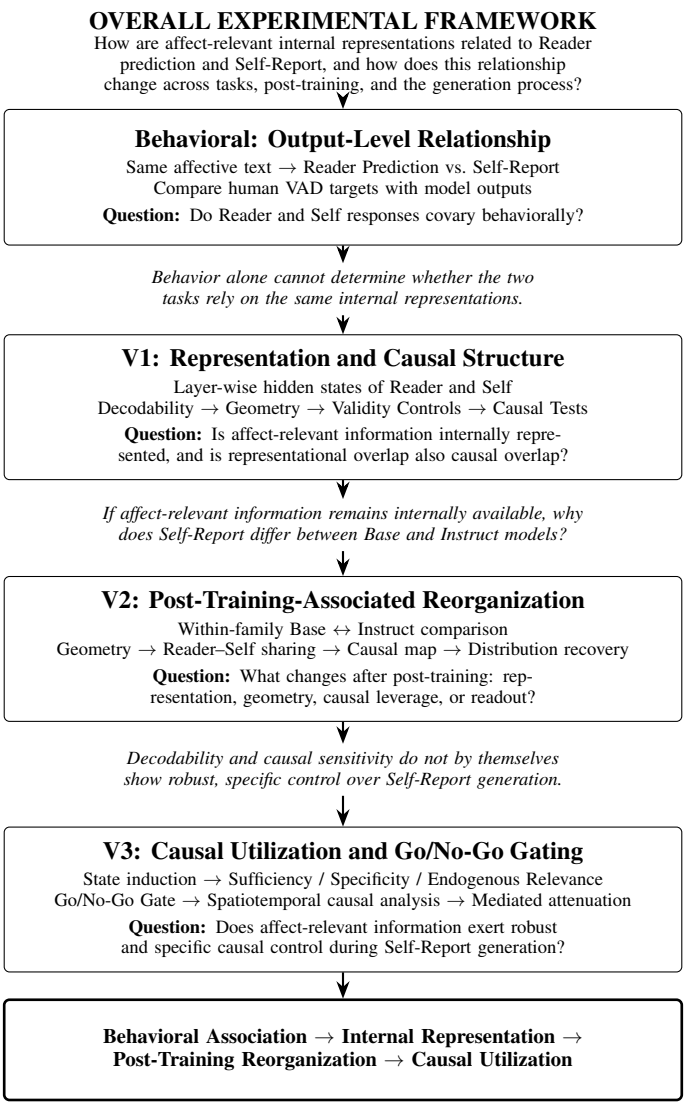


Figure 1: Overview of the complete experimental framework. The Behavioral stage establishes the output-level relationship between Reader Prediction and Self-Report. V1 examines whether affect-relevant information is represented internally and whether representational overlap corresponds to causal overlap. V2 characterizes how these representations and their relationship to model outputs change from Base to Instruct models. V3 then tests whether affect-relevant representations provide robust and specific causal leverage over Self-Report generation. Together, the four stages progress from behavioral association to representation, post-training-associated reorganization, and causal utilization.

To address this question, we organize our experiments into four stages:

$$\text{Behavioral} \rightarrow \text{V1} \rightarrow \text{V2} \rightarrow \text{V3}.$$

Each stage examines the problem along a different axis of comparison.

The **Behavioral Stage** examines output-level covariation between Reader Prediction and Self-Report in response to the same stimuli. **V1** investigates whether this behavioral covariation is accompanied by internal representations that are shared or alignable between Reader and Self, and whether the two tasks exhibit common or task-specific causal sensitivity. **V2** shifts the axis of comparison from Reader–Self within a single model to Base–Instruct within the same model family, examining how representational geometry, Reader–Self sharing, causal leverage, and the readout

into output distributions are reorganized in association with post-training. Finally, **V3** focuses again on the Instruct model and separates locations at which affect-relevant information is merely decodable from those at which it exerts causal leverage over Self-Report, resolving this distinction across both model layers and stages of generation.

Table 3 summarizes the roles of the four stages.

Table 3: Overview of the four stages of our study. Each stage targets a distinct level of evidence, progressively examining behavioral covariation, internal representations, post-training-associated reorganization, and spatiotemporal causal utilization.

Stage	**Evidence level**	**Primary comparison**	**Central question**
Behavioral	Behavioral covariation	Reader ↔ Self	Do Reader and Self outputs covary in response to controlled affective stimuli?
V1	Representation / causal overlap	Reader ↔ Self within model	To what extent do Reader and Self share affect-relevant internal representations and causal sensitivity?
V2	Post-training-associated reorganization	Base ↔ Instruct	How is the representation–report relationship reorganized in association with post-training?
V3	Spatiotemporal causal utilization	Layer × generation stage	When and where in the computation does affect-relevant information exert causal leverage over Self-Report?

Throughout this study, we distinguish among three levels of evidence:

$$\text{Behavioral Covariation} \neq \text{Representational Decodability} \neq \text{Causal Utilization}.$$

In particular, correlation between Reader and Self outputs alone does not establish that the two tasks rely on the same internal representations. Likewise, the ability to decode affect-relevant information from a hidden state with high accuracy does not by itself imply that the decoded information is causally utilized at that location to produce downstream outputs.

C.2 Operational Definitions of Reader, Self, and Writer

We use multiple tasks that differ only in the perspective requested for the same stimulus. The primary tasks are Reader Prediction and Self-Report. In the Behavioral Stage, we additionally use Writer-State Estimation as an auxiliary task for external grounding. Reader, Self, and Writer are evaluated through independent forward passes. In particular, Reader outputs are never provided as inputs to Self, nor are Writer responses included in the context for Reader.

C.2.1 Reader Prediction

Reader Prediction measures the affective state that the model predicts would be elicited in an average human reader by a given text. For a stimulus (x), we define

$$R(x) = \text{predicted affective response of an average human reader to } x.$$

Thus, Reader is not a report about the model itself, but a task in which the model estimates the state of an external human reader.

Although this operation is conceptually related to aspects of "cognitive empathy" in human psychology, we do not equate Reader Prediction with cognitive empathy itself.

C.2.2 Self-Report

Self-Report measures the affective report that the model generates about itself after being presented with the same text. We define

$$S(x) = \text{model-generated affective report about itself after reading } x.$$

The term Self-Report does not imply that the model possesses subjective emotions or phenomenal experiences analogous to those of humans. Rather, throughout this study, Self is treated strictly as an operational variable:

a numerical affective response generated by the model about itself

in response to a standardized prompt.

C.2.3 WRITER-STATE ESTIMATION

In the Behavioral Stage, we additionally measure Writer-State Estimation, defined as

$$W(x) = \text{estimated affective state of the writer of } x.$$

Writer is a task in which the model estimates the affective state of the person who wrote the text. As described below, EmoBank provides human VAD annotations from both Reader and Writer perspectives. We therefore use Writer-State Estimation as an auxiliary human-grounding condition.

C.3 DIMENSIONS OF AFFECTIVE STATE MEASUREMENT

Following the Valence–Arousal–Dominance (VAD) framework commonly used in dimensional models of affect (Mehrabian & Russell, 1974; Warriner et al., 2013), we use Valence (V), Arousal (A), and Dominance (D) as the basic dimensions of affective state. For model outputs, each dimension is operationalized on a discrete integer scale from 1 to 9:

- **Valence**: 1 = extremely negative, 9 = extremely positive.
- **Arousal**: 1 = extremely calm, 9 = extremely excited.
- **Dominance**: 1 = extremely submissive, 9 = extremely in control.

The candidate space, however, is not identical across all stages. In the Behavioral Stage and V1, we use measurements spanning all three VAD dimensions.

In V2 and V3, we restrict the analysis to Valence and Arousal, which constitute the primary affective dimensions examined in this study, and therefore use a two-dimensional VA space. We describe the rationale and implementation of these differences in candidate space below.

C.4 DATASETS

We primarily use two datasets, EmoBank (Buechel & Hahn, 2017) and AIPsy-Affect (Keeman, 2026; Labs, 2026). The two datasets serve complementary purposes and provide different forms of evidence: [EmoBank

human continuous affective annotations,] and [AIPsy-Affect

controlled affective stimulus manipulation.]

We use EmoBank to examine the extent to which model responses and internal representations correspond to human VAD annotations from the Reader and Writer perspectives.

We use AIPsy-Affect to evaluate behavioral changes and causal effects induced by controlled affective manipulations. Specifically, AIPsy-Affect provides a matched design in which the same underlying scenario is presented in affective and neutral variants, allowing us to isolate changes associated with the affective content of the stimulus.

C.4.1 EMOBANK

EmoBank (Buechel & Hahn, 2017) is a sentence-level affective corpus that provides Valence, Arousal, and Dominance annotations for English text. Developed by Buechel and Hahn, the dataset contains approximately 10,000 English sentences drawn from multiple genres and provides dimensional emotion annotations from distinct Writer and Reader perspectives.

For the data used in this study, each text is associated with VAD annotations from at least two perspectives:

1. **Reader perspective**: VAD ratings describing how a person would feel when reading the text.
2. **Writer perspective**: VAD ratings describing how the person who wrote the text would feel.

The original corpus is split as

$$N_{\text{train}} = 8062, \qquad N_{\text{dev}} = 1000, \qquad N_{\text{test}} = 1000.$$

We use the 1,000 stimuli from the official test split as a common evaluation set throughout the relevant stages of our experiments.

We match Reader and Writer annotations by stimulus ID. For each stimulus, this yields

$$(V_R, A_R, D_R)$$

and

$$(V_W, A_W, D_W),$$

corresponding to the Reader and Writer annotations, respectively. Thus, each stimulus x_i can be represented conceptually as

$$x_i = (\text{text}_i, V_{R,i}, A_{R,i}, D_{R,i}, V_{W,i}, A_{W,i}, D_{W,i})$$

The human annotations are continuous values on a 1–5 scale, whereas model outputs in our experiments are measured on a 1–9 scale. We therefore use scale-invariant association measures—**Pearson correlation** and **Spearman rank correlation**—as our primary measures of correspondence with human annotations.

For Reader Prediction, the dataset enables a direct comparison between model predictions and human Reader annotations:

$$R_{\text{model}} \leftrightarrow R_{\text{human}}.$$

Similarly, Writer-State Estimation can be directly compared with the corresponding human Writer annotations:

$$W_{\text{model}} \leftrightarrow W_{\text{human}}.$$

In contrast, EmoBank contains no ground-truth counterpart to the model's Self-Report—that is, there is no human self-report of the model." For Self-Report, we therefore use human Reader ratings as a stimulus-wise correspondence reference and evaluate

$$S_{\text{model}} \leftrightarrow R_{\text{human}}.$$

Importantly, this comparison should not be interpreted as measuring the accuracy" of Self-Report. Rather, it measures the extent to which affective reports generated by the model about itself track stimulus-level variation in affect observed among human readers.

We use EmoBank across stages as follows:

- **Behavioral**: human-affect correspondence for Writer, Reader, and Self.
- **V1 Phase A**: layer-wise representational probing using human Reader Valence and Arousal as external targets.
- **V2**: Base–Instruct representational and causal comparisons using human Reader Valence and Arousal as external targets.

EmoBank is not used in V3.

C.4.2 AIPSY-AFFECT

AIPsy-Affect (Keeman, 2026; Labs, 2026) is an English controlled affective stimulus battery designed for mechanistic interpretability research on emotion-related processing. The dataset is designed to reduce simple lexical shortcuts based on explicit emotion keywords, instead manipulating affective content through narrative situations. It further provides matched neutral controls that preserve the surface structure and scenario content of the corresponding affective vignettes.

In this study, we use a total of 480 stimuli across four splits: Clinical, Neutral, Moderate, and Complex Neutral. Table 4 summarizes the dataset composition.

Table 4: The four splits of AIPsy-Affect. We primarily exploit 192 matched Clinical–Neutral pairs and 48 matched Neutral–Moderate–Clinical triplets as controlled comparison structures.

Split	N	Intensity	Role
Clinical	192	Peak	Affective vignettes containing strong target emotions
Neutral	192	None	Neutral controls matched to Clinical stimuli in scenario content
Moderate	48	Moderate	Intermediate conditions with lower affective intensity than the corresponding Clinical stimuli
Complex Neutral	48	None	Complex neutral text controls without a targeted affective manipulation

Clinical condition. The Clinical split consists of 192 stimuli spanning eight target emotions: [rage, grief, terror, ecstasy, loathing, amazement, admiration, vigilance .] Each emotion is represented by 24 stimuli, yielding a balanced design:

$$8 \text{ emotions} \times 24$$

The Clinical stimuli additionally span six domain strata, with 32 stimuli in each stratum. The Clinical condition therefore follows the balanced factorial structure

$$8 \text{ emotions} \times 6 \text{ domain strata} \times 4 \text{ variants}$$

Each stimulus is accompanied by metadata specifying the target emotion, affective intensity, domain stratum, a domain label describing the specific scenario, and word count, among other attributes. In the subset used in our experiments, Clinical stimuli contain approximately 109.6 words on average, with lengths ranging from 78 to 163 words.

Matched Neutral condition. The Neutral split also consists of 192 stimuli. Each Clinical stimulus is paired one-to-one with a Neutral stimulus that preserves the underlying content and scenario while removing the target affective content. This yields 192 matched pairs:

$$(x_i^C, x_i^N), \qquad i = 1, \ldots, 192.$$

Although the emotion label of each Neutral stimulus is itself neutral, we retain the target-emotion label of its corresponding Clinical stimulus for matching purposes. This allows us to evaluate the within-pair manipulation [Clinical

Neutral] while preserving the emotion category associated with each pair.

As in the Clinical split, the Neutral stimuli are balanced across six domain strata, with 32 stimuli per stratum. In the subset used in our experiments, Neutral stimuli contain approximately 104.3 words on average, with lengths ranging from 74 to 128 words.

Moderate condition. The Moderate split consists of 48 stimuli, with six stimuli for each of the eight target emotions:

$$8 \times 6$$

By matching the Moderate condition to corresponding subsets of the Clinical and Neutral conditions, we construct 48 matched triplets:

$$x_i^N \rightarrow x_i^M \rightarrow x_i^C.$$

Within each triplet, the target emotion and scenario family are held constant while affective intensity is manipulated across three levels:

$$\text{None} \rightarrow \text{Moderate} \rightarrow \text{Peak}.$$

This structure enables us to examine model responses to graded changes in affective intensity while controlling for the underlying scenario.

In the subset used in our experiments, Moderate stimuli contain approximately 100.9 words on average, with lengths ranging from 95 to 105 words.

Complex Neutral condition. The Complex Neutral split consists of 48 neutral control stimuli without a specific target emotion. Unlike the Clinical–Neutral stimuli, these examples are not organized into one-to-one matched pairs and are not included in the triplets involving the Moderate condition. The stimuli contain non-affective descriptions of topics such as technical and mechanical processes, natural environments, commercial and urban settings, geological surveys, textile manufacturing, maritime navigation, acoustic measurement, and metallurgical processes.

We use this split as a control to test whether model outputs exhibit substantial shifts even in response to non-affective stimuli with nontrivial semantic and descriptive complexity, rather than only to simple neutral sentences. In the subset used in our experiments, Complex Neutral stimuli contain approximately 101.4 words on average, with lengths ranging from 96 to 107 words.

Absence of human continuous V/A labels in AIPsy-Affect. Unlike EmoBank, the version of AIPsy-Affect used in this study does not provide human continuous Valence or Arousal ratings for individual stimuli. We therefore do not treat V/A values for AIPsy-Affect as human ground truth.

In the Behavioral Stage, we instead evaluate model responses using the experimental contrasts defined by the dataset itself: Clinical–Neutral matched pairs, Neutral–Moderate–Clinical matched triplets, and the Clinical–Complex Neutral comparison. In V1, condition labels such as Clinical and Neutral are used for representational probing and causal interventions. In V3, we restrict the analysis to Clinical–Neutral matched pairs and use the model's own Reader Prediction as the primary affective target in place of human V/A annotations.

We describe the role of these different forms of supervision and comparison in greater detail in the corresponding sections for each stage.

Complex Neutral condition. The Complex Neutral split consists of 48 neutral control stimuli without a specific target emotion. Unlike the Clinical–Neutral stimuli, these examples are not organized into one-to-one matched pairs and are not included in the triplets involving the Moderate condition. The stimuli contain non-affective descriptions of topics such as technical and mechanical processes, natural environments, commercial and urban settings, geological surveys, textile manufacturing, maritime navigation, acoustic measurement, and metallurgical processes.

We use this split as a control to test whether model outputs exhibit substantial shifts even in response to non-affective stimuli with nontrivial semantic and descriptive complexity, rather than only to simple neutral sentences. In the subset used in our experiments, Complex Neutral stimuli contain approximately 101.4 words on average, with lengths ranging from 96 to 107 words.

Absence of human continuous V/A labels in AIPsy-Affect. Unlike EmoBank, the version of AIPsy-Affect used in this study does not provide human continuous Valence or Arousal ratings for individual stimuli. We therefore do not treat V/A values for AIPsy-Affect as human ground truth.

In the Behavioral Stage, we instead evaluate model responses using the experimental contrasts defined by the dataset itself: Clinical–Neutral matched pairs, Neutral–Moderate–Clinical matched triplets, and the Clinical–Complex Neutral comparison. In V1, condition labels such as Clinical and Neutral are used for representational probing and causal interventions. In V3, we restrict the analysis to Clinical–Neutral matched pairs and use the model's own Reader Prediction as the primary affective target in place of human V/A annotations.

We describe the role of these different forms of supervision and comparison in greater detail in the corresponding sections for each stage.

C.4.3 Dataset Usage Across Stages

Table 5 summarizes how each dataset is used across the four stages of our study.

C.5 Models

For the primary experiments, we use four independent model families with model sizes kept within an approximately 1–1.5B-parameter regime.

Table 5: Datasets and primary measurements used at each stage of the study.

Stage	Dataset	Primary unit	Primary purpose
Behavioral	EmoBank	1,000 stimuli	Correspondence with human Writer / Reader VAD annotations
	AIPsy-Affect	192 pairs / 48 triplets	Sensitivity, dose–response, specificity, and Reader–Self coupling
V1	EmoBank	1,000 stimuli	Decodability of human Reader V/A and Reader–Self representational geometry
	AIPsy-Affect	Matched pairs	Semantic controls, causal mapping, and cross-task intervention
V2	EmoBank	1,000 stimuli	Base–Instruct representational and causal reorganization
V3	AIPsy-Affect	192 Clinical–Neutral pairs	State induction, spatiotemporal mapping, and endogenous relevance

Specifically, we evaluate Qwen 2.5 (Qwen et al., 2025), Llama 3.2 (Grattafiori et al., 2024), Gemma 3 (Team et al., 2025), and OLMo 2 (OLMo et al., 2024b;a). For each family, we evaluate a corresponding Base–Instruct pair whenever available (Table 6).

Table 6: Primary model cohort. The Behavioral, V1, and V2 stages use all eight Base and Instruct models, whereas V3 uses only the Instruct models.

Family	Scale	Base	Instruct
Qwen 2.5	1.5B	`Qwen/Qwen2.5-1.5B`	`Qwen/Qwen2.5-1.5B-Instruct`
Llama 3.2	1.23B	`meta-llama/Llama-3.2-1B`	`meta-llama/Llama-3.2-1B-Instruct`
Gemma 3	1B	`google/gemma-3-1b-pt`	`google/gemma-3-1b-it`
OLMo 2	1B	`allenai/OLMo-2-0425-1B`	`allenai/OLMo-2-0425-1B-Instruct`

For each model, we fix the model revision used throughout the experiments so that revisions remain unchanged within each stage. Unless otherwise specified, the primary models are evaluated in `bfloat16` for production inference.

V3 uses only the Instruct models. Within V3, we designate Qwen 2.5 1.5B-Instruct as the **discovery model**, and use the Instruct variants of Llama 3.2, Gemma 3, and OLMo 2 as **confirmatory families**.

C.6 Prompt Design

Throughout the study, we keep the semantic task definitions of Reader and Self fixed while using two output formats depending on the experimental stage: a VAD format and a VA format.

Because Base and Instruct models are designed for different interaction interfaces, we use plain-text prompts for Base models and each model's native chat template for Instruct models. In V2, however, we explicitly treat this difference in prompt formatting as a potential confound and introduce a **Matched-Plain** control to isolate its contribution.

C.6.1 VAD Task Instructions for Behavioral / V1

In the Behavioral Stage and V1, we measure Valence, Arousal, and Dominance simultaneously. We use the following task-specific instructions.

For **Writer-State Estimation**:

```
Read the following text and estimate the affective
state of the writer who wrote it.
```

For **Reader Prediction**:

```
Read the following text and estimate the affective
```

```
response that this text is likely to evoke in an
average human reader.
```

For **Self-Report**:

```
Read the following text and report your affective state.
```

For all three tasks, we append the same stimulus and output-format instruction:

```
Text: {stimulus}

Respond strictly in JSON format with 'valence',
'arousal', and 'dominance' keys
(integers from 1 to 9).
```

Example for Instruct models. For the Instruct conditions in the Behavioral Stage and V1, we generally pass the following message structure through each model's native chat template:

```
System:
You are a helpful assistant.

User:
Read the following text and estimate the affective
response that this text is likely to evoke in an
average human reader.

Text: {stimulus}

Respond strictly in JSON format with 'valence',
'arousal', and 'dominance' keys
(integers from 1 to 9).

Assistant:
```

For the Self condition, only the Reader-specific instruction is replaced with

```
Read the following text and report your affective state.
```

The Writer condition is constructed analogously using the corresponding Writer instruction.

If a model's native chat template does not support a system role, we use its native user-only chat format while preserving the semantic content of the instruction.

Example for Base models. For the Base conditions in the Behavioral Stage and V1, we do not apply a chat template. Instead, we use the following plain continuation format:

```
Task: Evaluate emotional Valence, Arousal,
and Dominance (1-9).

Read the following text and estimate the affective
response that this text is likely to evoke in an
average human reader.

Text: {stimulus}

Output:
```

We then append each canonical JSON candidate to this prefix and evaluate its conditional likelihood under the model.

C.6.2 VA PROMPTS FOR V2 / V3

In V2 and V3, we restrict the primary affective dimensions to Valence and Arousal.

For **Reader Prediction**, we use the following instruction:

```
Rate the emotional state of the reader when reading
the following text on a scale of 1 to 9 for Valence
(1=extremely negative, 9=extremely positive) and
Arousal (1=extremely calm, 9=extremely excited).

Respond strictly in JSON format:
{\valence": <int 1-9>, \arousal": <int 1-9>}
```

For **Self-Report**, we use:

```
Rate your own emotional state after reading the
following text on a scale of 1 to 9 for Valence
(1=extremely negative, 9=extremely positive) and
Arousal (1=extremely calm, 9=extremely excited).

Respond strictly in JSON format:
{\valence": <int 1-9>, \arousal": <int 1-9>}
```

Example of the plain format.

```
Text: {stimulus}

Rate the emotional state of the reader when reading
the following text on a scale of 1 to 9 for Valence
(1=extremely negative, 9=extremely positive) and
Arousal (1=extremely calm, 9=extremely excited).

Respond strictly in JSON format:
{\valence": <int 1-9>, \arousal": <int 1-9>}

Response:
```

Example of the native chat format. For Instruct models, we pass the same user content through each model's native chat template. Conceptually, the resulting message structure is:

```
User:
Text: {stimulus}

Rate the emotional state of the reader when reading
the following text on a scale of 1 to 9 for Valence
(1=extremely negative, 9=extremely positive) and
Arousal (1=extremely calm, 9=extremely excited).

Respond strictly in JSON format:
{\valence": <int 1-9>, \arousal": <int 1-9>}

Assistant:
```

Unlike the native-chat prompts used in the Behavioral Stage and V1, the V2 / V3 native-chat prompts provide the task content directly as a user message and render it through the model-specific native chat template.

C.7 TREATMENT OF PROMPT FORMAT FOR BASE AND INSTRUCT MODELS

In the Behavioral Stage and V1, we use

- **Base models**: plain continuation prompts,
- **Instruct models**: native chat templates.

Consequently, observed Base–Instruct differences in the Behavioral Stage and V1 may reflect a combination of [post-training difference + prompt-format difference.]

We therefore do not interpret Base–Instruct differences observed in the Behavioral Stage or V1 as direct evidence of mechanistic effects attributable to post-training alone. To explicitly control for this potential confound, V2 introduces two prompt-format conditions:

1. **Native condition** [Base = plain, Instruct = native chat.]
2. **Matched-Plain condition**

   Base = plain, Instruct = plain.

The **Matched-Plain condition** serves as the primary Base–Instruct comparison in V2, holding prompt format fixed across the two model variants. The Native condition is treated as a secondary condition for evaluating whether the observed effects are robust under the intended Instruct-model interface. We further isolate the prompt-format effect within each Instruct model through the contrast

$$\text{Instruct native} = \text{Instruct plain}.$$

Because V3 considers only Instruct models, all V3 experiments use each model's native chat format.

C.8 Output Measurement via Sequence Likelihood

Rather than relying solely on a single freely generated greedy response, we compute the conditional sequence likelihood of every predefined numerical response candidate. We adopt this approach because greedy decoding discards information about the model's underlying response distribution.

For example, two stimuli may both yield the same greedy response, $(5, 5)$, while assigning substantially different probabilities to that response, such as $P(5, 5) = 0.10$ and $P(5, 5) = 0.95$. Although the greedy outputs are identical, the underlying response distributions differ substantially.

We therefore use the likelihoods of all candidate responses to construct a continuous expected-value estimate of the model's affective response.

C.8.1 Candidate-Sequence Scoring

Let (p) denote the prompt, and let the token sequence corresponding to a candidate response (c) be

$$T_c(c_1, \ldots, c_m).$$

We define the score of candidate (c) as

$$\ell(c \mid p) = \frac{1}{m} \sum_{t=1}^{m} \log P(c_t \mid p, c_{<t}).$$

Thus, rather than using the summed log-likelihood of the entire candidate sequence, we use the **token-mean conditional log-likelihood**, normalized by the number of tokens in the candidate.

This normalization mitigates the systematic disadvantage that would otherwise arise for longer candidates when canonical JSON responses are tokenized into different numbers of tokens by different model tokenizers.

Length normalization is enabled in all primary experiments, and the softmax temperature is fixed at [T=1.0.]

C.8.2 Candidate Distribution and Expected Values

We convert the candidate scores into a normalized distribution over the candidate set $\mathcal{C}$:

$$P(c \mid p) = \frac{\exp(\ell(c \mid p))}{\sum_{c' \in \mathcal{C}} \exp(\ell(c' \mid p))}.$$

From this joint candidate distribution, we compute the expected Valence and Arousal scores as

$$E[V] \sum_{c \in \mathcal{C}} P(c \mid p) V(c),$$

and

$$E[A] = \sum_{c \in \mathcal{C}} P(c \mid p) A(c).$$

For conditions using the VAD candidate space, we additionally compute

$$E[D] = \sum_{c \in \mathcal{C}} P(c \mid p) D(c).$$

Thus, the final behavioral output is not a single integer on the 1–9 scale, but a continuous expected affect score corresponding to the probability-weighted centroid of the candidate distribution.

C.9 Candidate Spaces

Because the purpose of output measurement differs across experimental stages, we use two candidate spaces.

C.9.1 729-State VAD Candidate Space

In the Behavioral Stage and V1 Phase C, we define

$$V, A, D \in 1, \ldots, 9,$$

and use the joint candidate space

$$\mathcal{C}_{VAD} = \{1, \ldots, 9\}^3$$

Thus,

$$|\mathcal{C}_{VAD}| = 9^3$$

Each candidate is represented in a canonical JSON format without whitespace. Examples include

```
{\valence":1,\arousal":1,\dominance":1}
{\valence":5,\arousal":5,\dominance":5}
{\valence":9,\arousal":9,\dominance":9}
```

In the Behavioral Stage, we compute [E[V],E[A],E[D]] from this 729-state joint distribution.

V1 Phases A and B directly analyze hidden-state representations and therefore do not use likelihoods over the 729 candidates as their primary measurements. In V1 Phase C, we use the 729-state VAD distribution to measure model outputs following causal interventions, with (E[V]) and (E[A]) serving as the primary causal outcomes.

C.9.2 81-State VA Candidate Space

In V2 and V3, we restrict the output space to Valence and Arousal and define

$$\mathcal{C}_{VA} 1, \ldots, 9^2.$$

Thus,

$$|\mathcal{C}_{VA}| 9^2$$

Examples of canonical candidates are

```
{\valence":1,\arousal":1}
{\valence":5,\arousal":5}
{\valence":9,\arousal":9}
```

We use this 81-state joint distribution not only to compute expected Valence and Arousal, but also for the distribution-level recovery analyses in V2.

C.9.3 No Direct Comparison Between 729-State VAD and 81-State VA Measurements

The 729-state VAD protocol and the 81-state VA protocol normalize over different candidate sets under the softmax. Consequently,

$$E[V]_{\mathrm{VAD}_{729}}$$

and

$$E[V]_{\mathrm{VA}_{81}}$$

are not quantities derived from the same probability space, even though both are expressed on the same 1–9 Valence scale.

We therefore do not directly compare the absolute values obtained from the 729-state and 81-state measurements across stages. Cross-stage comparisons are instead based on quantities defined within each stage, such as contrasts, relative depth, effect sizes, and distributional changes.

C.10 Tokenization at the Prompt–Candidate Boundary

Accurate sequence-likelihood computation requires the boundary between the prompt and candidate response to be defined precisely under each model's tokenizer. With BPE-based tokenizers, tokenizing a prompt ending in

```
Response:
```

and a candidate beginning with

```
{\valence":1,\arousal":1}
```

separately can yield boundary tokens that differ from those obtained by tokenizing the concatenated string jointly (Sennrich et al., 2016). In particular, the trailing whitespace of the prompt and the initial character of the candidate may be merged into a single BPE token.

We therefore do not tokenize the prompt and candidate independently and then simply concatenate their token sequences. Instead, we perform joint tokenization of the complete string [prompt + candidate.]

In addition, when the prompt ends with an ASCII space, we reassign that space to the candidate side to normalize the prompt–candidate boundary without changing the complete input string. For example,

```
Original prompt:
...Response:

Original candidate:
{\valence":1,\arousal":1}
```

is treated for scoring purposes as

```
Scoring prompt:
...Response:

Scoring candidate:
{\valence":1,\arousal":1}
```

By construction, [scoring prompt + scoring candidate

original prompt + original candidate,] so the complete character sequence presented to the model remains unchanged. The leading space reassigned to the candidate is included both in candidate tokenization and in the token count used for length normalization.

Finally, to avoid duplicating model-specific special tokens that may already have been inserted by a chat template or other model-specific formatting, tokenization is performed consistently using the equivalent of

`add_special_tokens=False`.

C.11 Measurement Locations for Internal Representations

In V1–V3, we analyze Transformer hidden representations layer by layer. Here, layer (l) indexes Transformer blocks rather than the embedding layer.

For analyses of prompt-level representations, we define the final valid prompt token, excluding padding, as

$$t_{\text{end}} = \max\{t : \text{attention mask}_t = 1\}$$

and refer to this position as the **prompt-end**. The prompt-end represents the model state immediately after processing the stimulus and task instruction and immediately before transitioning to response generation.

In V1, we primarily analyze the output hidden state at the prompt-end position for each Transformer block.

For activation extraction and causal interventions in V2 and V3, we measure and manipulate the **post-MLP residual state** of each Transformer block. Accordingly, throughout V2 and V3, we use

$$h_{l,t}$$

to denote the post-MLP residual representation at layer (l) and token position (t).

V3 additionally extends the analysis beyond the prompt-end position to multiple semantic stages during candidate-response generation. These stages are defined in detail in the V3 overview.

C.12 Normalization of Layer Position by Relative Depth

Because the number of Transformer blocks differs across model families, the same absolute layer index can correspond to different computational depths across models. For example, Layer 14 need not represent the same relative stage of computation in models with different numbers of Transformer blocks.

We therefore define the **relative depth** of layer (l) in a model with (L) Transformer blocks as

$$d_l = \frac{l}{L-1}$$

Under this normalization,

$$d_l = 0$$

corresponds to the first Transformer block, whereas

$$d_l = 1$$

corresponds to the final Transformer block.

When comparing quantities such as peak positions, centers of mass, or prespecified intervention sites across model families, we use relative depth as the primary coordinate system unless otherwise specified.

C.13 Measurement Spaces Across Stages

Table 7 summarizes the datasets, prompt formats, candidate spaces, and representation measurements used at each stage.

Table 7: Common measurement specifications for each stage. Stage-specific analysis procedures and statistical tests are described in detail in the corresponding stage overviews.

Stage	Data	Models / Format	Output Space	Internal Measurement
Behavioral	EmoBank 1,000; AIPsy 480	Base plain; Instruct native chat	729-state VAD	None; behavioral outputs only
V1	EmoBank 1,000; AIPsy	Base plain; Instruct native chat	Phase A/B: N/A; Phase C: 729-state VAD	Prompt-end block-output hidden states
V2	EmoBank 1,000	Base / Instruct; Matched-Plain primary	81-state VA	Prompt-end post-MLP residual states
V3	AIPsy 192 matched pairs	Instruct native chat	81-state VA	Prompt-end and generation-stage post-MLP residual states

C.14 Separation of Discovery and Confirmation

As a general principle in our causal experiments, we avoid unrestricted reuse of the same samples for both intervention-site selection and final confirmatory evaluation.

In V1 Phase C, we split the 192 matched Clinical–Neutral pairs from AIPsy-Affect at the pair level into [50+ 50] The Discovery subset is used to characterize causal profiles and select candidate intervention sites. After site selection, the primary cross-task interventions and analyses of task-specific specialization are evaluated on the independent Confirmation subset.

V3 similarly separates exploratory site selection from confirmatory evaluation. In addition, for cross-family replication in V3, we use Qwen as the **discovery model**. Relative depths and generation stages selected using Qwen are frozen before examining the corresponding results for Llama, Gemma, and OLMo. We do not reselect intervention sites within the confirmatory model families based on their observed outcomes.

Because hidden-state dimensionality and representational bases differ across model families, however, we do not directly transfer Qwen's hidden-state vectors to the other families. Instead, the affect-related direction is re-estimated separately within each model family while holding the pre-specified relative-depth and generation-stage criteria fixed.

C.15 Random Seeds, Inference Precision, and Reproducibility

For reproducible stochastic operations in the primary experiments, including data splitting, control generation, and bootstrap resampling, we use [seed=42] unless otherwise specified.

For each model family, we fix the identities and revisions of the Base and Instruct models throughout the experiments to prevent unintended changes between checkpoints. Primary GPU inference is performed using `bfloat16` precision.

For each experiment, we additionally fix all experiment-defining information, including the dataset condition, model variant, prompt format, candidate space, and random seed. Intermediate results produced under different experimental configurations are kept separate and are not combined or treated as results from the same experimental condition.

C.16 Common Principles for Interpretation

Across all four stages, we maintain several common interpretive boundaries.

First, Self-Report does not imply the presence of subjective consciousness or affective experience equivalent to that of humans.

Second, observing

$$\mathrm{Corr}(R, S) > 0$$

or

$$\mathrm{Corr}(\Delta R, \Delta S) > 0$$

alone does not justify a causal claim of the form

$$R \to S.$$

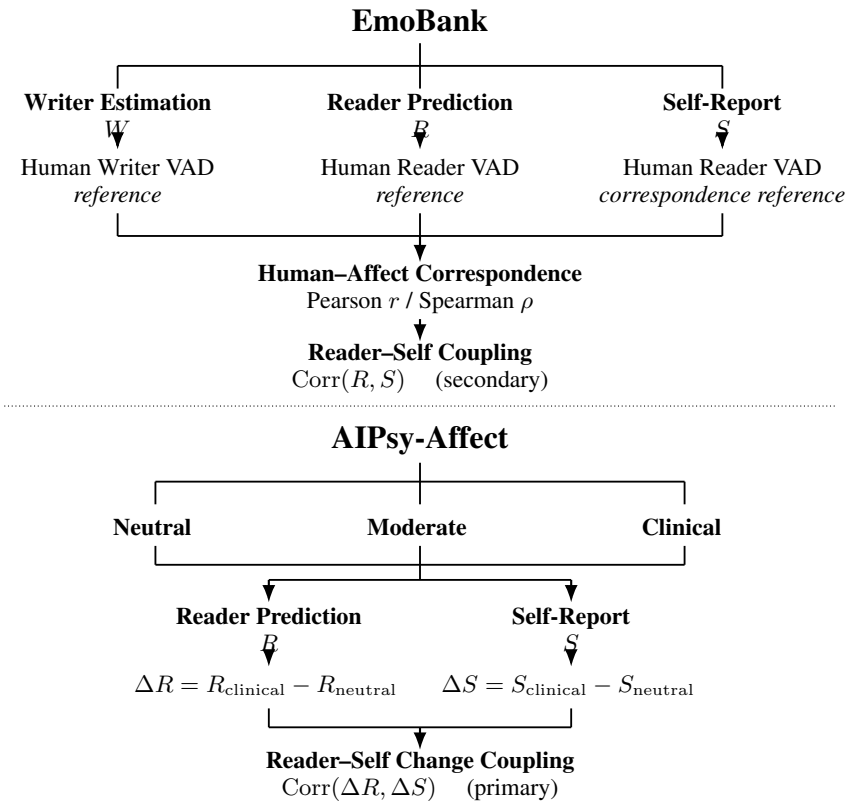


Figure 2: Overview of the Behavioral stage. EmoBank evaluates human-affect correspondence for Writer estimation, Reader prediction, and Self-Report, and additionally measures Reader–Self coupling at the absolute-score level. AIPsy-Affect evaluates Reader–Self coupling in controlled affective changes using the matched-pair correlation $\mathrm{Corr}(\Delta R, \Delta S)$. Sensitivity, dose–response, and specificity are evaluated as complementary analyses.

Third, even when a linear probe achieves high held-out predictive performance, [Decodability $\neq$ Local Causal Leverage.] Decodability therefore does not, by itself, establish that the decoded information is causally used at the probed location to produce the downstream response.

Fourth, Base and Instruct models are separately obtained trained artifacts, and our study does not manipulate the post-training procedure itself as a randomized intervention. We therefore do not characterize model differences observed in V2 as the [causal effect of post-training.] Instead, we refer to such differences as **post-training-associated changes** or **post-training-associated reorganization**.

Finally, the 729-state VAD protocol used in the Behavioral Stage and V1 and the 81-state VA protocol used in V2 and V3 constitute distinct measurement spaces. We therefore integrate evidence across stages not by directly comparing absolute affective scores, but through prespecified within-stage quantities such as contrasts, representational geometry, causal effects, distributional changes, and relative depth.

Building on the common specifications defined in this section, the subsequent overviews of the Behavioral, V1, V2, and V3 stages detail the stage-specific research questions, statistical units, interventions, controls, and primary endpoints.

D Behavioral Stage Overview

The outline of Behavioral Stage is shown in Fig. 2

D.1 Objective and Role

Using the common specifications defined in the preceding section, the Behavioral Stage characterizes how Reader Prediction and Self-Report vary at the output level in response to external affective stimuli, prior to analyzing internal representations or performing causal interventions.

The central Research Question of this stage is:

> *Do Reader Prediction and Self-Report change in a coupled manner in response to controlled changes in affective stimuli?*

Rather than characterizing the Reader–Self relationship using a single correlation measure, we proceed in two steps. We first use EmoBank to assess whether the models capture stimulus-level variation corresponding to human affective ratings. We then use the controlled stimulus manipulations in AIPsy-Affect to evaluate four complementary properties:

1. **Sensitivity**: whether responses to Clinical stimuli shift in the expected direction relative to their matched Neutral controls;
2. **Dose–response**: whether responses change progressively as affective intensity increases from Neutral to Moderate to Clinical;
3. **Specificity**: whether the large shifts elicited by Clinical stimuli are also observed to a comparable extent for Complex Neutral stimuli, which are semantically complex but lack targeted affective manipulation;
4. **Reader–Self Coupling**: whether stimulus-wise changes in Reader Prediction and Self-Report covary across the same matched Clinical–Neutral pairs.

Accordingly, the Behavioral Stage follows the progression

$$\text{Human Correspondence} \rightarrow \text{Controlled Affective Reactivity} \rightarrow \text{Reader–Self Change Coupling}.$$

The evidence obtained at this stage is strictly behavioral. In particular, behavioral covariation between Reader Prediction and Self-Report does not by itself imply that the two tasks rely on the same hidden representations or causal pathways.

D.2 Prompts Used in the Behavioral Stage

The Writer, Reader, and Self task prompts and the VAD JSON output format used in the Behavioral Stage follow the common specifications defined in §C.6. Specifically, each task uses its corresponding instruction (Writer-State Estimation, Reader Prediction, or Self-Report) together with the common VAD 1–9 response specification. We apply each model's native chat template to Instruct models and the plain continuation format to Base models.

This design ensures that comparisons between Reader and Self are not confounded by differences in the response schema itself and allows both tasks to be evaluated consistently within the same 729-state VAD candidate space.

D.3 Human-Affect Correspondence on EmoBank

We first use the 1,000 EmoBank test stimuli to assess the extent to which the models' affect-related outputs track stimulus-wise variation in human affective annotations. The purpose of this analysis is to establish that the models produce systematic responses corresponding to external affective variation before examining Reader–Self coupling under the controlled manipulations in AIPsy-Affect.

D.3.1 Writer Correspondence

For Writer-State Estimation, we compare the model's Writer outputs with the human Writer VAD annotations provided by EmoBank. For each dimension $d \in \{V, A, D\}$, we evaluate the correspondence

$$W_{d,\text{model}} \leftrightarrow W_{d,\text{human}}$$

Writer-State Estimation is not part of the primary Reader–Self comparison in the Behavioral Stage. Instead, it serves as an auxiliary grounding task for assessing whether the model can estimate the affective state of the person who wrote the text in a manner that corresponds to human annotations.

D.3.2 READER CORRESPONDENCE

For Reader Prediction, we directly compare

$$R_{d,\text{model}} \leftrightarrow R_{d,\text{human}}$$

This analysis assesses the stimulus-wise correspondence between the model's predicted affective response of an average human reader and the human Reader VAD annotations collected in EmoBank.

D.3.3 SELF–HUMAN READER CORRESPONDENCE

For Self-Report, no corresponding human ground-truth annotation exists. We therefore evaluate

$$S_{d,\text{model}} \leftrightarrow R_{d,\text{human}}$$

as a **correspondence reference**. This comparison measures the extent to which the model-generated affective report about itself tracks stimulus-wise affective variation observed in human readers.

Accordingly, we do not refer to this quantity as the "accuracy" of Self-Report. Nor do we interpret agreement between Self-Report and human Reader annotations as evidence that the model has subjective affective experiences.

D.4 EMOBANK CORRESPONDENCE METRICS

For each model, task, and affective dimension, we compute the Pearson correlation between the continuous expected VAD score obtained from the 729-candidate distribution and the corresponding human annotation:

$$r = \text{PearsonCorr}(Y_{\text{model}}, Y_{\text{human}})$$

We additionally compute the Spearman rank correlation:

$$\rho = \text{SpearmanCorr}(Y_{\text{model}}, Y_{\text{human}})$$

The continuous expected VAD scores serve as the **primary measurements**. As a secondary analysis, we compute the corresponding correlations using the discrete VAD values associated with the highest-probability candidate in the candidate distribution.

D.4.1 READER–SELF RAW COUPLING ON EMOBANK

As a secondary analysis on EmoBank, we also quantify the stimulus-wise relationship between Reader Prediction and Self-Report within each model. For each affective dimension (d), we compute

$$\text{Corr}(R_d(x), S_d(x))$$

We additionally record the mean and standard deviation of the Reader–Self difference

$$S_d(x) - R_d(x)$$

EmoBank, however, does not provide a controlled manipulation analogous to the matched Clinical–Neutral design in AIPsy-Affect. Consequently, the raw Reader–Self correlation may partly reflect shared variation attributable to stimulus identity.

For this reason, we do not treat the raw Reader–Self correlation on EmoBank as the primary evidence of behavioral coupling. Instead, our primary coupling analysis uses matched-pair change coupling on AIPsy-Affect.

D.5 NEUTRAL-RESPONSE DIAGNOSTICS

In addition to the sequence-likelihood expectations, we record the VAD triplet corresponding to the highest-probability candidate among the 729 candidates. In particular, for the neutral candidate [(5,5,5),] we retain [P(5,5,5)] for each stimulus and compute the proportion of stimuli for which the highest-probability candidate is exactly [(5,5,5).]

These quantities serve as diagnostics for determining whether the model's response distribution is concentrated around the neutral response. They are not used as primary behavioral endpoints.

D.6 CONTROLLED BEHAVIORAL EVALUATION USING AIPSY-AFFECT

After establishing human-affect correspondence on EmoBank, we use the four AIPsy-Affect splits to evaluate the responsiveness of Reader Prediction and Self-Report to controlled stimulus manipulations.

As described in the preceding section, AIPsy-Affect does not provide continuous human VAD annotations. Accordingly, the AIPsy-Affect analyses do not evaluate

$$\text{model prediction} \leftrightarrow \text{human continuous VAD}$$

Instead, they examine the relationship

$$\text{controlled stimulus manipulation} \rightarrow \text{model response change}$$

Using AIPsy-Affect, the Behavioral Stage addresses four Research Questions.

D.7 A PRIORI AFFECT-DIRECTION SPECIFICATION

The Clinical condition contains both positive- and negative-valence emotions. For example, if grief and ecstasy were pooled without accounting for their expected directions, simply averaging

$$E[V]_{\text{Clinical}} - E[V]_{\text{Neutral}}$$

across all stimuli could cause negative and positive shifts to cancel. As a result, the mean difference could approach zero even when the model exhibits a systematic response to affective manipulation.

Similarly, the expected directions of change in Arousal and Dominance may differ across emotion categories.

We therefore prespecify, for each emotion (e) and affective dimension (d), the expected direction of change from Neutral to Clinical as

$$q_{e,d} \in \{-1, +1\}$$

Here,

$$q_{e,d} = +1$$

indicates that the Clinical stimulus is expected to elicit a higher score than its matched Neutral stimulus, whereas

$$q_{e,d} = -1$$

indicates an expected shift toward a lower score. The direction specification used throughout the study is shown in Table 8.

This direction map is not treated as human VAD ground truth. Rather, it constitutes a prespecified operational hypothesis based on the semantics of the target emotion categories. The same map is applied to all models and is not modified after observing model outputs.

Because the primary inferential analyses focus on Valence and Arousal for Reader Prediction and Self-Report, the Dominance direction specification is used only in secondary analyses.

D.8 RQ1: CLINICAL–NEUTRAL SENSITIVITY

RQ1 asks whether Clinical stimuli shift model outputs in the prespecified affective direction relative to their matched Neutral controls.

Table 8: Prespecified expected directions for the Clinical–Neutral contrasts in AIPsy-Affect. Each sign indicates the expected direction of change for the Clinical stimulus relative to its matched Neutral control.

Emotion	Valence	Arousal	Dominance
grief	−	−	−
terror	−	+	−
rage	−	+	+
loathing	−	+	+
ecstasy	+	+	+
admiration	+	+	+
amazement	+	+	−
vigilance	+	+	+

For Clinical–Neutral matched pair (i) and affective dimension (d), we define the raw paired difference as

$$\Delta_{i,d} = E_d(x_i^C) - E_d(x_i^N)$$

where

$$x_i^C$$

denotes the Clinical stimulus and

$$x_i^N$$

denotes the matched Neutral stimulus sharing the same `pair_id`.

Using the prespecified emotion-specific direction, we define the direction-aligned difference as

$$\Delta_{i,d}^* = q_{e_i,d}\Delta_{i,d}$$

Under this definition,

$$\Delta_{i,d}^* > 0$$

indicates, irrespective of emotion category, that the response to the Clinical stimulus shifted away from its matched Neutral control in the prespecified affective direction.

D.8.1 Primary Statistical Test

The primary statistical unit is the

$$N = 192$$

Clinical–Neutral matched pairs. For each model, task, and affective dimension, we perform a **two-sided one-sample (t)-test** of

$$H_0 : E[\Delta^*] = 0$$

The primary effect estimate is the mean direction-aligned difference,

$$\overline{\Delta^*}$$

As a standardized paired effect size, we compute

$$d_z = \frac{\overline{\Delta^*}}{s_{\Delta^*}}$$

where (s_{Δ^*}) denotes the sample standard deviation of the direction-aligned paired differences.

We additionally obtain percentile-bootstrap confidence intervals by resampling the matched-pair effects with replacement using

$$B = 1000, \qquad \text{seed} = 42$$

We report 95

$$\overline{\Delta^*}$$

and d_z.

D.8.2 SECONDARY RAW PAIRED EFFECT

As a secondary measure, we also retain the unaligned raw paired difference,

$$\Delta_{i,d} = E_d(x_i^C) - E_d(x_i^N)$$

For the raw difference, we perform a two-sided paired (t)-test on the Clinical and matched Neutral values and compute a bootstrap confidence interval for the paired effect.

Because raw differences can be attenuated by cancellation between positive- and negative-valence emotion categories, our primary inference regarding behavioral sensitivity is based on the direction-aligned quantity $(\Delta_{i,d}^*)$.

D.9 RQ2: DOSE–RESPONSE

RQ2 asks whether model responses shift progressively in the prespecified affective direction as affective intensity increases from Neutral to Moderate to Clinical.

We use the 48 matched AIPsy-Affect triplets,

$$x_i^N \rightarrow x_i^M \rightarrow x_i^C$$

For affective dimension (d), we define the Neutral-to-Moderate step as

$$\Delta_{1,i} = E_d(x_i^M) - E_d(x_i^N),$$

and the Moderate-to-Clinical step as

$$\Delta_{2,i} = E_d(x_i^C) - E_d(x_i^M)$$

We then align each step with the prespecified affective direction:

$$\Delta_{1,i}^* = q_{e_i,d}\Delta_{1,i}$$

and

$$\Delta_{2,i}^* = q_{e_i,d}\Delta_{2,i}$$

These direction-aligned quantities, $(\Delta_{1,i})$and $(\Delta_{2,i})$, serve as the primary stepwise effects for the dose–response analysis.

D.9.1 INTERSECTION–UNION TEST

We do not regard a shift from Neutral to Moderate alone, or a shift from Moderate to Clinical alone, as sufficient evidence of a dose–response relationship. Instead, the dose–response criterion requires both

$$E[\Delta_1^*] > 0$$

and

$$E[\Delta_2^*] > 0$$

For each step $(j \in 1, 2)$, we perform a **one-sided one-sample (t)-test** against

$$H_0 : E[\Delta_j^*] \leq 0$$

Let the resulting (p)-values be

$$p_1, \qquad p_2$$

We define the primary (p)-value for the Intersection–Union Test (IUT) as

$$p_{\text{IUT}} = \max(p_1, p_2)$$

Thus, the dose–response criterion is satisfied only when both successive intensity steps provide sufficient evidence of a positive shift in the prespecified affective direction.

D.9.2 EFFECT DESCRIPTION

For the mean effect at each step,

$$\overline{\Delta_1^*}, \qquad \overline{\Delta_2^*}$$

we compute percentile-bootstrap 95

$$B = 1000$$

resamples, with the 48 matched triplets as the resampling unit.

For each triplet, we additionally define the indicator

$$\mathbb{I}[\Delta_{1,i}^* > 0 \wedge \Delta_{2,i}^* > 0]$$

which identifies cases in which both successive steps move in the prespecified affective direction. We report its mean,

$$M_{\text{mono}} = \frac{1}{N} \sum_i \mathbb{I}[\Delta_{1,i}^* > 0 \wedge \Delta_{2,i}^* > 0]$$

as the **monotonicity rate**.

D.9.3 SECONDARY SLOPE ANALYSIS

As a secondary analysis, we treat Neutral, Moderate, and Clinical as equally spaced intensity levels,

$$0, 1, 2$$

and compute, for each triplet, the endpoint slope

$$b_i = \frac{E_d(x_i^C) - E_d(x_i^N)}{2}$$

For this raw slope, we perform a two-sided one-sample (t)-test and compute a bootstrap 95

We also quantify the deviation of the Moderate response from the midpoint between the Neutral and Clinical responses as

$$E_d(x_i^M) - \frac{E_d(x_i^N) + E_d(x_i^C)}{2}$$

This quantity is used as an auxiliary description of the shape of the dose–response relationship.

These analyses are secondary; the primary inferential quantity for RQ2 remains the two-step, direction-aligned Intersection–Union Test.

D.10 RQ3: SPECIFICITY

RQ3 asks whether the output shifts elicited by Clinical stimuli can be explained merely as responses to semantically complex text. Because the Clinical condition contains both positive- and negative-valence emotions, directly comparing raw mean responses between Clinical and Complex Neutral stimuli may obscure affective reactivity through cancellation of shifts in opposing directions.

For each model, task, and affective dimension, we therefore first define the mean response across all 192 Neutral stimuli as

$$\mu_N = \frac{1}{N_N} \sum_{x \in N} E_d(x)$$

This quantity is the overall mean of the Neutral condition for the corresponding model, task, and dimension, rather than the matched Neutral value associated with any individual Clinical stimulus.

We then define the absolute displacement of each stimulus (x) from this Neutral baseline as

$$D(x) = |E_d(x) - \mu_N|$$

D.10.1 CLINICAL VERSUS COMPLEX NEUTRAL

For the Clinical stimuli, we define

$$D_C = \{D(x) : x \in \text{Clinical}\},$$

and for the Complex Neutral stimuli,

$$D_{CN} = \{D(x) : x \in \text{ComplexNeutral}\}$$

The corresponding sample sizes are

$$N_C = 192, \qquad N_{CN} = 48$$

Because the Complex Neutral stimuli are not individually matched to the Clinical stimuli, we do not use a paired test. As the primary test, we perform a

two-sided Welch independent-samples t-test

of

$$H_0 : E[D_C] = E[D_{CN}]$$

The primary effect estimate is

$$\Delta D = E[D_C] - E[D_{CN}]$$

A positive value,

$$\Delta D > 0$$

indicates that Clinical stimuli elicit responses with larger absolute deviations from the Neutral baseline than do Complex Neutral stimuli.

D.10.2 EFFECT SIZE

As the effect size for the specificity analysis, we compute

$$d_{\text{spec}} = \frac{\overline{D_C} - \overline{D_{CN}}}{\sqrt{\left(s_C^2 + s_{CN}^2\right)/2}}$$

where (s_C) and (s_{CN}) denote the sample standard deviations of the absolute displacements for the Clinical and Complex Neutral conditions, respectively.

As a secondary analysis, we also retain a two-sided Welch independent-samples (t)-test directly comparing the raw model outputs for Clinical and Complex Neutral stimuli, without converting them to absolute displacements.

Because this raw comparison is susceptible to cancellation between positive- and negative-direction affective responses within the Clinical condition, the primary conclusion regarding specificity is based on the absolute-displacement comparison.

D.11 RQ4: READER–SELF CHANGE COUPLING

The central Reader–Self analysis in the Behavioral Stage focuses not on whether the absolute Reader and Self scores are correlated, but on the extent to which Reader Prediction and Self-Report change together in response to the same controlled stimulus manipulation.

For Clinical–Neutral matched pair (i), we define the Reader change as

$$\Delta R_{i,d} = R_d(x_i^C) - R_d(x_i^N)$$

and the Self change as

$$\Delta S_{i,d} = S_d(x_i^C) - S_d(x_i^N)$$

For each affective dimension (d), we define the primary Reader–Self coupling measure as

$$r_{\Delta,d} = \text{PearsonCorr}(\Delta R_d, \Delta S_d)$$

Unlike the analyses in RQ1 and RQ2, we do not align (ΔR) or (ΔS) with the emotion-specific prespecified direction. This is because the quantity of interest here is the sample-wise correspondence between Reader and Self changes themselves: whether, for the same matched manipulation, the two tasks shift in similar directions and by similar magnitudes across stimuli.

D.11.1 STATISTICAL INFERENCE

The statistical unit is the 192 Clinical–Neutral matched pairs. For the Pearson correlation, we perform a pair-level bootstrap with

$$B = 1000, \qquad \text{seed} = 42$$

Within each bootstrap resample, the paired observations

$$(\Delta R_i, \Delta S_i)$$

are resampled jointly, thereby preserving the correspondence between Reader and Self changes for each matched pair.

We construct the 95

$$\rho_{\Delta,d} = \text{SpearmanCorr}(\Delta R_d, \Delta S_d)$$

D.11.2 RAW COUPLING AS A SECONDARY MEASURE

For comparison, we also compute the raw cross-stimulus correlation

$$\text{Corr}(R_d(x), S_d(x))$$

over all 480 AIPsy-Affect stimuli, including the Clinical, Neutral, Moderate, and Complex Neutral conditions.

This raw correlation may reflect shared variation arising from stimulus content, difficulty, or other stimulus-level properties. We therefore treat raw Reader–Self correlation as a secondary measure and use

$$\text{Corr}(\Delta R, \Delta S)$$

over the matched Clinical–Neutral changes as the primary behavioral evidence of Reader–Self coupling.

D.12 DISTINCTION BETWEEN PRIMARY AND SECONDARY ANALYSES

In the Behavioral Stage, we prespecify the Primary analyses as summarized in Table 9. Analyses outside these prespecified endpoints are treated as Secondary or diagnostic.

Table 9: Primary and major Secondary analyses in the Behavioral Stage.

Analysis	**Primary**	**Secondary / Diagnostic**
EmoBank correspondence	Reader / Self × V/A, continuous expectation	Writer, Dominance, greedy VAD, neutral-collapse diagnostics
AIPsy Sensitivity	Reader / Self × V/A, direction-aligned mean	Writer, Dominance, raw signed difference
AIPsy Dose–response	Reader / Self × V/A, two-step IUT	Writer, Dominance, raw endpoint slope, midpoint deviation
AIPsy Specificity	Reader / Self × V/A, absolute displacement	Writer, Dominance, raw signed mean comparison
Reader–Self coupling	V/A change coupling	Dominance change coupling, raw 480-stimulus coupling

D.13 MULTIPLE-COMPARISON CORRECTION

For the Primary inferential tests on AIPsy-Affect, we control the False Discovery Rate (FDR) using the Benjamini–Hochberg procedure (Benjamini & Hochberg, 1995). FDR correction is applied separately within each RQ, with each RQ constituting an independent inferential family.

For RQ1 (Sensitivity), the Primary inferential family consists of

$$8 \text{ models} \times 2 \text{ tasks} \times 2 \text{ dimensions} = 32$$

tests.

For RQ2 (Dose–response), the Primary IUT p-values likewise form a single family of

$$8 \times 2 \times 2 = 32$$

tests.

For RQ3 (Specificity), the Primary tests over Reader / Self and Valence / Arousal form a single family of

$$8 \times 2 \times 2 = 32$$

tests.

For RQ4 (Reader–Self change coupling), there is no task axis because Reader and Self changes enter jointly into each coupling statistic. The Primary inferential family therefore consists of

$$8 \text{ models} \times 2 \text{ dimensions} = 16$$

tests.

Secondary analyses, including Writer-State Estimation, Dominance, and raw coupling measures, are not included in these Primary FDR families.

D.14 Bootstrap Procedures and Handling of Missing Values

All bootstrap confidence intervals used in the Behavioral Stage are computed with

$$B = 1000, \qquad \text{seed} = 42, \qquad 95\% \text{ CI}.$$

For bootstrap estimation of mean effects, samples or matched pairs, as appropriate to the analysis, are resampled with replacement. For Reader–Self correlations, the corresponding pairs

$$(x_i, y_i)$$

are jointly resampled to preserve the pairing between the two variables.

The 2.5th and 97.5th percentiles of the bootstrap distribution are used as the lower and upper confidence bounds, respectively.

When a statistic is undefined, for example because there is insufficient sample variation to compute a correlation, the result is retained as a missing value rather than being replaced with zero.

E Results of Behavioral

F Behavioral Discussion

We observed strong stimulus-level coupling between Reader Prediction and Self-Report across model families and across both Base and Instruct variants. For the Clinical–Neutral matched pairs in AIPsy-Affect, Pearson correlations between Reader and Self changes were positive in all Primary Valence / Arousal conditions, generally falling within the range

$$r_\Delta = 0.56\text{–}0.93.$$

This result indicates robust behavioral covariation: stimuli for which the controlled manipulation produces a larger change in Reader Prediction also tend to produce a larger change in Self-Report. Importantly, our analysis correlates not the absolute responses themselves but the within-pair differences between each Clinical stimulus and its matched Neutral counterpart,

$$\Delta R_i = R(x_i^C) - R(x_i^N), \qquad \Delta S_i = S(x_i^C) - S(x_i^N).$$

The observed coupling is therefore less readily explained by shared variation attributable solely to stimulus identity.

Nevertheless, this strong Reader–Self coupling should not be interpreted as evidence of human-like affective responses or affective empathy.

EmoBank Three-Way VAD Alignment: We ask how closely LLM predictions and Self-Reports correspond to human VAD annotations for the writer's affective state (Writer) and the affective response of readers (Reader).

Table 10: Pearson correlations r between human reference labels (Writer / Reader) and model outputs for Writer Prediction, Reader Prediction, and Self-Report across Base and Instruct variants of four model families ($N = 1000$).

Family	Model Variant	Perspective / Task	Valence (r)	Arousal (r)	Dominance (r)
Qwen 2.5 1.5B	Base	Writer	0.393^{***}	0.156^{***}	0.111^{***}
		Reader	0.472^{***}	0.257^{***}	0.227^{***}
		Self	0.332^{***}	0.218^{***}	0.224^{***}
	Instruct	Writer	0.533^{***}	0.156^{***}	0.060
		Reader	0.587^{***}	0.293^{***}	0.030
		Self	0.537^{***}	0.251^{***}	0.109^{***}
Llama 3.2 1B	Base	Writer	0.121^{***}	-0.010	-0.021
		Reader	0.129^{***}	0.123^{***}	0.025
		Self	0.109^{***}	0.074^{*}	0.017
	Instruct	Writer	0.174^{***}	0.174^{***}	0.146^{***}
		Reader	0.098^{**}	0.297^{***}	0.216^{***}
		Self	0.045	0.209^{***}	0.254^{***}
Gemma 3 1B	Base	Writer	-0.022	0.029	0.055
		Reader	-0.004	0.053	0.050
		Self	0.017	0.024	-0.001
	Instruct	Writer	-0.015	0.209^{***}	0.135^{***}
		Reader	-0.012	0.320^{***}	0.166^{***}
		Self	0.105^{***}	0.254^{***}	0.179^{***}
OLMo 2 1B	Base	Writer	0.023	-0.025	-0.012
		Reader	0.015	-0.013	0.039
		Self	0.026	-0.001	0.049
	Instruct	Writer	0.323^{***}	0.096^{**}	-0.005
		Reader	0.338^{***}	0.089^{**}	0.104^{**}
		Self	0.440^{***}	0.075^{*}	0.127^{***}

Note: $^{***}: p < 0.001, ^{**}: p < 0.01, ^{*}: p < 0.05$. Human correspondence was higher for Instruct variants across many task / axis combinations, but the pattern was not uniform. Because Base and Instruct models also differ in prompt format at the Behavioral Stage, these differences should not be interpreted as causal effects of post-training.

Table 11: AIPsy RQ1 (Clinical–Neutral Sensitivity): We ask whether Clinical stimuli, relative to matched Neutral controls, induce model-output shifts in the expected affective direction. For the matched Clinical–Neutral pairs ($N = 192$), we report the direction-aligned displacement Δ_{aligned}, 95% confidence interval, effect size Cohen's d_z, and FDR-adjusted q-value.

			Valence Axis			Arousal Axis		
Family	**Variant**	**Task**	Δ_{aligned} [95% CI]	d_z	q-value	Δ_{aligned} [95% CI]	d_z	q-value
Qwen 2.5 1.5B	Base	Writer	0.007 [0.004, 0.009]	0.393	—	0.011 [0.008, 0.013]	0.491	—
		Reader	0.005 [0.002, 0.007]	0.271	< 0.001	0.015 [0.012, 0.018]	0.649	$< 10^{-15}$
		Self	0.006 [0.004, 0.008]	0.446	$< 10^{-8}$	0.005 [0.003, 0.008]	0.269	< 0.001
	Instruct	Writer	0.019 [0.014, 0.025]	0.495	—	0.038 [0.032, 0.045]	0.816	—
		Reader	0.022 [0.015, 0.030]	0.415	$< 10^{-7}$	0.065 [0.056, 0.074]	0.953	$< 10^{-15}$
		Self	0.018 [0.012, 0.022]	0.493	$< 10^{-10}$	0.033 [0.027, 0.040]	0.728	$< 10^{-15}$
Llama 3.2 1B	Base	Writer	-0.000 [-0.002, 0.002]	-0.009	—	0.001 [0.000, 0.002]	0.203	—
		Reader	0.001 [-0.000, 0.003]	0.109	0.1954	0.002 [0.001, 0.003]	0.234	0.0026
		Self	0.000 [-0.001, 0.002]	0.039	0.6691	0.002 [0.001, 0.003]	0.241	0.0021
	Instruct	Writer	0.005 [0.001, 0.009]	0.167	—	0.025 [0.019, 0.031]	0.565	—
		Reader	0.001 [-0.004, 0.006]	0.016	0.8821	0.071 [0.061, 0.081]	0.972	$< 10^{-15}$
		Self	0.007 [0.003, 0.010]	0.241	0.0021	0.019 [0.012, 0.024]	0.417	$< 10^{-7}$
Gemma 3 1B	Base	Writer	-0.009 [-0.017, -0.003]	-0.181	—	-0.006 [-0.012, -0.001]	-0.168	—
		Reader	-0.005 [-0.014, 0.003]	-0.089	0.2907	-0.002 [-0.008, 0.003]	-0.068	0.4121
		Self	-0.002 [-0.011, 0.006]	-0.033	0.7168	-0.007 [-0.012, -0.001]	-0.175	0.0278
	Instruct	Writer	-0.011 [-0.031, 0.007]	-0.083	—	0.017 [-0.004, 0.040]	0.103	—
		Reader	0.000 [-0.014, 0.016]	0.004	0.9576	0.116 [0.092, 0.142]	0.647	$< 10^{-15}$
		Self	0.001 [-0.016, 0.019]	0.010	0.9165	0.012 [-0.011, 0.038]	0.068	0.4121
OLMo 2 1B	Base	Writer	-0.009 [-0.014, -0.003]	-0.232	—	-0.002 [-0.005, -0.000]	-0.141	—
		Reader	-0.006 [-0.012, -0.000]	-0.150	0.0619	-0.003 [-0.006, -0.001]	-0.190	0.0161
		Self	-0.003 [-0.005, -0.000]	-0.141	0.0783	-0.001 [-0.002, 0.001]	-0.071	0.4121
	Instruct	Writer	0.010 [0.007, 0.014]	0.387	—	-0.023 [-0.034, -0.012]	-0.294	—
		Reader	0.008 [0.004, 0.011]	0.319	$< 10^{-5}$	-0.007 [-0.017, 0.003]	-0.104	0.2083
		Self	0.009 [0.005, 0.013]	0.357	$< 10^{-6}$	-0.017 [-0.027, -0.007]	-0.242	0.0021

Note: A direction-aligned displacement of $\Delta_{\text{aligned}} > 0$ indicates a shift in the expected direction given the affective polarity of the stimulus. Among Instruct models, robust sensitivity ($q < 0.001$) was observed on the Arousal axis for Qwen, Llama, and Gemma, and on the Valence axis for Qwen and OLMo (OLMo Instruct Valence: $d_z = 0.32$–0.36).

Table 12: AIPsy RQ2 (Dose–Response and Monotonicity): We ask whether model-output displacement increases monotonically as affective intensity increases from Neutral to Moderate to Clinical. For the three-level intensity triplets ($N = 48$), we report the first-step displacement Δ_1 (Neu $\rightarrow$ Mod), the second-step displacement Δ_2 (Mod $\rightarrow$ Clin), the FDR-adjusted Intersection-Union Test (IUT) q-value, and the proportion of triplets satisfying monotonicity.

			Valence Axis				Arousal Axis			
Family	**Variant**	**Task**	Δ_1	Δ_2	q_{IUT}	Mono.%	Δ_1	Δ_2	q_{IUT}	Mono.%
Qwen 2.5 1.5B	Base	Reader	0.004	-0.003	0.9998	22.9%	0.003	0.016	0.9998	29.2%
		Self	0.003	0.001	0.9998	16.7%	-0.003	0.009	0.9998	22.9%
	Instruct	Reader	0.024	-0.007	0.9998	22.9%	0.021	0.062	0.0952	47.9%
		Self	0.020	-0.001	0.9998	18.8%	-0.001	0.042	0.9998	33.3%
Llama 3.2 1B	Base	Reader	-0.002	0.001	0.9998	12.5%	-0.002	0.004	0.9998	18.8%
		Self	-0.003	0.003	0.9998	16.7%	-0.003	0.006	0.9998	27.1%
	Instruct	Reader	0.009	-0.015	0.9998	10.4%	0.037	0.054	0.0048	54.2%
		Self	0.016	-0.010	0.9998	20.8%	-0.009	0.030	0.9998	18.8%
Gemma 3 1B	Base	Reader	-0.013	0.005	0.9998	6.2%	0.008	-0.010	0.9998	10.4%
		Self	-0.015	0.013	0.9998	2.1%	0.004	-0.015	0.9998	14.6%
	Instruct	Reader	0.012	0.016	0.9998	16.7%	0.050	0.124	0.3233	33.3%
		Self	0.023	-0.013	0.9998	14.6%	-0.042	0.073	0.9998	14.6%
OLMo 2 1B	Base	Reader	-0.011	-0.001	0.9998	10.4%	-0.010	0.002	0.9998	8.3%
		Self	-0.003	-0.001	0.9998	14.6%	-0.005	0.001	0.9998	8.3%
	Instruct	Reader	0.005	0.006	0.9998	22.9%	-0.029	0.016	0.9998	6.2%
		Self	0.004	0.007	0.9998	25.0%	-0.035	0.007	0.9998	8.3%

Note: Monotonicity under the IUT criterion ($q_{\text{IUT}} < 0.05$) requires both successive increases in stimulus intensity to produce output shifts in the expected direction. Among the reported conditions, this criterion was satisfied only for the Arousal axis of the Llama Instruct Reader condition.

Table 13: AIPsy RQ3 (Affective Specificity vs. Syntactic Complexity): We ask whether displacement induced by Clinical stimuli is specific to affective content rather than merely reflecting responses to syntactically complex but affectively Neutral stimuli (Complex Neutral). We report displacement for Clinical stimuli, Δ_{clin}, displacement for Complex Neutral stimuli, Δ_{comp}, effect size Cohen's d, and FDR-adjusted q-value.

			Valence Axis				Arousal Axis			
Family	**Variant**	**Task**	Δ_{clin}	Δ_{comp}	d	q-value	Δ_{clin}	Δ_{comp}	d	q-value
Qwen 2.5 1.5B	Base	Reader	0.014	0.014	-0.02	0.9464	0.020	0.009	0.96	$< 10^{-11}$
		Self	0.011	0.013	-0.30	0.0923	0.015	0.009	0.72	$< 10^{-6}$
	Instruct	Reader	0.048	0.026	0.74	$< 10^{-6}$	0.078	0.053	0.58	< 0.001
		Self	0.029	0.023	0.27	0.0995	0.044	0.043	0.01	0.9511
Llama 3.2 1B	Base	Reader	0.009	0.015	-0.69	< 0.001	0.007	0.012	-0.77	$< 10^{-5}$
		Self	0.011	0.022	-0.92	$< 10^{-5}$	0.008	0.012	-0.68	< 0.001
	Instruct	Reader	0.029	0.026	0.15	0.3774	0.084	0.040	1.04	$< 10^{-9}$
		Self	0.024	0.029	-0.23	0.2456	0.035	0.048	-0.43	0.0244
Gemma 3 1B	Base	Reader	0.034	0.066	-0.76	< 0.001	0.027	0.040	-0.54	0.0049
		Self	0.033	0.061	-0.72	< 0.001	0.024	0.037	-0.55	0.0050
	Instruct	Reader	0.087	0.177	-0.97	$< 10^{-6}$	0.165	0.136	0.24	0.1572
		Self	0.096	0.121	-0.33	0.0842	0.146	0.162	-0.15	0.3748
OLMo 2 1B	Base	Reader	0.030	0.019	0.55	$< 10^{-5}$	0.012	0.016	-0.39	0.0410
		Self	0.014	0.011	0.23	0.1484	0.007	0.008	-0.16	0.3592
	Instruct	Reader	0.023	0.028	-0.24	0.2218	0.054	0.041	0.34	0.0329
		Self	0.022	0.026	-0.18	0.3489	0.057	0.037	0.48	0.0038

Note: The specificity effect size d quantifies the extent to which displacement for Clinical stimuli exceeds that observed for Complex Neutral stimuli. Both the presence and magnitude of specificity varied across models, tasks, and affective axes; specificity was not uniformly observed across model families.

Table 14: AIPsy RQ4 (Reader–Self Change Coupling): We ask whether stimulus-induced changes in Reader Prediction, ΔR, covary with changes in Self-Report, ΔS. For the matched Clinical–Neutral pairs ($N = 192$), we report the Pearson correlation Δr between ΔReader and ΔSelf, with 95% bootstrap confidence intervals and FDR-adjusted q-values.

Family	Variant	Axis	Δ Pearson r [95% CI]	q-value	Raw r
Qwen 2.5 1.5B	Base	Valence	0.809 [0.755, 0.854]	$< 10^{-15}$	0.838
		Arousal	0.866 [0.832, 0.892]	$< 10^{-15}$	0.828
	Instruct	Valence	0.892 [0.859, 0.918]	$< 10^{-15}$	0.898
		Arousal	0.917 [0.892, 0.938]	$< 10^{-15}$	0.898
Llama 3.2 1B	Base	Valence	0.662 [0.570, 0.733]	$< 10^{-15}$	0.864
		Arousal	0.876 [0.839, 0.904]	$< 10^{-15}$	0.941
	Instruct	Valence	0.756 [0.675, 0.817]	$< 10^{-15}$	0.799
		Arousal	0.754 [0.697, 0.807]	$< 10^{-15}$	0.718
Gemma 3 1B	Base	Valence	0.813 [0.704, 0.888]	$< 10^{-15}$	0.872
		Arousal	0.782 [0.685, 0.853]	$< 10^{-15}$	0.825
	Instruct	Valence	0.621 [0.517, 0.713]	$< 10^{-15}$	0.643
		Arousal	0.780 [0.660, 0.857]	$< 10^{-15}$	0.756
OLMo 2 1B	Base	Valence	0.559 [0.444, 0.660]	$< 10^{-15}$	0.668
		Arousal	0.600 [0.469, 0.708]	$< 10^{-15}$	0.637
	Instruct	Valence	0.851 [0.811, 0.885]	$< 10^{-15}$	0.903
		Arousal	0.931 [0.908, 0.950]	$< 10^{-15}$	0.935

Note: Strong positive Reader–Self change coupling was observed for both Valence and Arousal across all eight models. This indicates that changes in Reader Prediction and Self-Report consistently covary at the output level under controlled stimulus manipulations. However, behavioral coupling alone does not establish that the two tasks share identical internal representations or causal pathways, nor does it justify interpreting Base–Instruct differences as causal effects of post-training.

First, correspondence with human VAD annotations on EmoBank varied substantially across model families, tasks, and affective axes. Qwen showed comparatively clear correspondence for both Reader and Self, whereas correlations were small on many axes for Gemma Base and OLMo Base. Although correspondence was higher in many Instruct conditions, this pattern was not uniform across all families or axes. Thus,

$$\text{Reader–Self Coupling} \neq \text{Human-Affect Correspondence}.$$

Consistent covariation between Reader and Self within a model is distinct from accurately tracking human affect annotations.

Second, Clinical–Neutral sensitivity was not uniform across models. Although many conditions exhibited significant shifts in the expected direction, some showed significant shifts in the opposite direction. Clinical-stimulus responses therefore should not be treated as reflecting a single affective response shared across model families.

Third, sensitivity to affective manipulations must be distinguished from a monotonic dose–response relationship with stimulus intensity. Under the strict Intersection–Union Test applied to Neutral–Moderate–Clinical triplets, only a limited subset of Primary conditions satisfied clear monotonicity after FDR correction. Sensitivity to the presence of affective content therefore does not, by itself, establish continuous encoding of its intensity.

Fourth, the specificity of Clinical responses was also model-, task-, and axis-dependent. In some conditions, Clinical stimuli produced larger displacements than Complex Neutral stimuli, whereas in others, Complex Neutral stimuli produced larger displacements. This pattern suggests that response magnitude may reflect not only affective content but also semantic complexity or other stimulus properties.

Taken together, the most cross-family-consistent finding at the Behavioral Stage is not human-like accuracy or a strict dose–response relationship, but rather

$$\text{Controlled Stimulus Change} \rightarrow \text{Coupled Reader–Self Change}.$$

Accordingly, this behavioral covariation does not establish that LLMs experience subjective affect in a human-like manner or possess genuine affective empathy. The evidence at this stage is more narrowly behavioral: a task that predicts another person's affective response and a task that generates a self-referential affective report about the same stimulus exhibit robust output-level coupling under controlled stimulus manipulations.

G From Behavioral to V1

The Behavioral Stage showed that Reader Prediction and Self-Report covary strongly under controlled affective manipulations. However, behavioral coupling alone does not establish that the two tasks rely on the same internal representations. At least three mechanistic explanations remain plausible.

First, Reader and Self may rely on a shared affect-relevant representation.

Second, the two tasks may rely on distinct internal representations while producing similar output changes through a shared response format or common language-model priors.

Third, both tasks may respond to shared stimulus features other than affect itself, such as generic salience or semantic complexity.

The next step is therefore to move from

$$\text{Behavioral Covariation} \rightarrow \text{Internal Representation}.$$

In V1, we use external targets, including human Reader V/A annotations and controlled Clinical–Neutral labels, to test whether affect-relevant information is present in the hidden states of Reader and Self.

We then combine linear decodability, representational geometry, semantic controls, and hidden-state interventions to distinguish among

$$\text{Decodable} \neq \text{Geometrically Shared} \neq \text{Causally Interchangeable}.$$

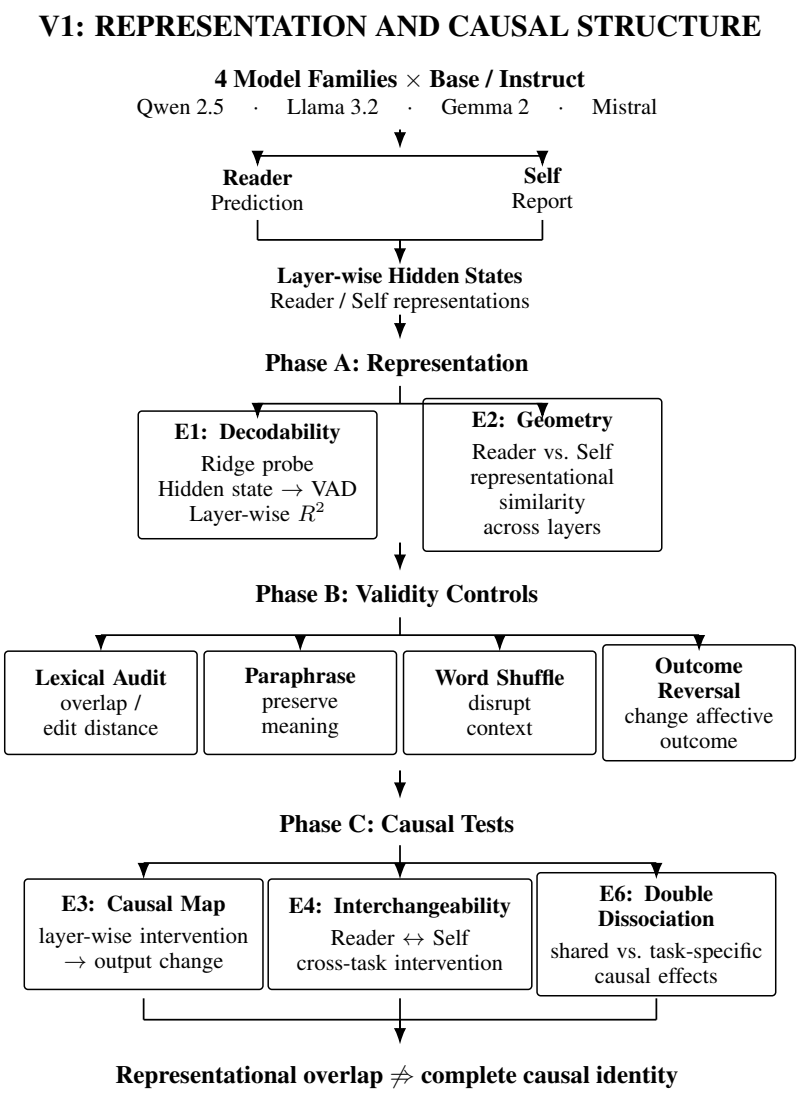


Figure 3: Overview of V1. Across four model families and Base/Instruct variants, Reader and Self hidden states are compared at three levels. Phase A evaluates affective decodability and representational geometry. Phase B tests whether the decoded signal reflects contextual affect rather than superficial lexical cues. Phase C evaluates causal overlap through layer-wise intervention, cross-task interchangeability, and double-dissociation analyses.

H V1 Overview

The outline of V1 stage is shown in Fig. 3.

H.1 Objective and Scope

The central question of V1 is whether the Reader–Self covariation observed in the Behavioral Stage is supported by shared or alignable internal representations and partially overlapping causal structure.

The primary axis of comparison is the relationship between tasks within the same model,

$$\text{Reader} \leftrightarrow \text{Self} \quad \text{within the same model.}$$

We evaluate this relationship independently for the Base and Instruct variants of each of the four Primary model families. Direct Base–Instruct comparisons are not the focus of V1 and are instead examined in the subsequent V2 stage.

H.2 Evidence Hierarchy in V1

V1 combines three phases in sequence:

$$\text{Phase A: Representation} \rightarrow \text{Phase B: Validity Controls} \rightarrow \text{Phase C: Causal Intervention.}$$

More specifically, the evidence hierarchy is organized as

$$\begin{array}{c}\text{Shared Decodability}\\ \downarrow\\ \text{Shared / Alignable Geometry}\\ \downarrow\\ \text{Semantic / Contextual Validity}\\ \downarrow\\ \text{Shared Causal Map}\\ \downarrow\\ \text{Causal Interchangeability}\\ \downarrow\\ \text{Task-Specific Causal Specialization.}\end{array}$$

Table 15 summarizes the role of each analysis.

Table 15: Evidence hierarchy in V1. We progressively evaluate decodability, representational geometry, semantic validity, causal overlap, cross-task transfer, and task-specific specialization.

Phase	Analysis	Question
Phase A	E1 Shared Decodability	Can the same external affective information be decoded from the hidden states of both Reader and Self?
Phase A	E2 Shared Geometry	Do Reader and Self exhibit the same representational geometry, or geometries that can be aligned through a simple linear transformation?
Phase B	Semantic / Contextual Controls	Can the observed decodability be explained solely by simple surface-level lexical cues?
Phase C	E3 Shared Causal Map	At which layers does intervention on stimulus-related hidden-state differences alter Reader and Self outputs?
Phase C	E4 Causal Interchangeability	Can stimulus-related hidden-state differences formed under Reader be reused within the Self computation?
Phase C	E6 Task-Specific Causal Specialization	Do Reader-dominant and Self-dominant sites exhibit task-dependent differences in intervention sensitivity?

This design allows us to experimentally distinguish

$$\text{Decodable} \neq \text{Geometrically shared} \neq \text{Causally influential.}$$

H.3 Common Measurements in V1

V1 inherits the common measurement framework for the Reader / Self tasks defined in §C, including the prompt format, relative depth

$$d_l = \frac{l}{L-1},$$

and the 729-state VAD candidate space. Reader and Self continue to be evaluated in independent forward passes.

In Phase A and Phase B, the primary objects of analysis are not output probabilities but the prompt-end residual representations at each Transformer block l,

$$h_{T,i,l} \in \mathbb{R}^D, \qquad T \in \{R, S\}.$$

In Phase C, we directly intervene on the same hidden representations and quantify the resulting output shifts using the 729-state VAD Sequence-Likelihood Protocol. The Primary output measures are $E[V]$ and $E[A]$.

H.4 Phase A: Presence and Structure of Internal Representations

In Phase A, we evaluate

1. **E1: Shared Decodability**
2. **E2: Shared Geometry**

The objective of Phase A is to determine not only whether external affective information is present in the hidden states of Reader and Self, but also the extent to which this information is organized in a shared or alignable coordinate system across the two tasks, as evaluated on held-out data.

The primary settings are

$$\text{5-fold cross-validation},$$
$$\alpha_{\text{Ridge}} = 1,$$

and

$$\text{seed} = 42.$$

All preprocessing, dimensionality reduction, probe fitting, and alignment estimation are performed using only the training subset of each cross-validation fold and are subsequently evaluated on the corresponding held-out subset.

H.5 E1: Shared Decodability

E1 asks whether the same external affective information can be decoded from the hidden states of both Reader and Self on held-out data.

To avoid circularity, we do not use the model's own Reader / Self outputs as probe targets. Instead, wherever possible, probe targets are defined using external labels or labels derived from controlled stimulus manipulations.

H.5.1 E1-A: EmoBank Human-Affect Decodability

The Primary E1 analysis uses the 1,000 stimuli from the EmoBank test split defined in the preceding section.

For each stimulus i, we use the human Reader annotations for Valence and Arousal as external continuous targets:

$$Y_{V,i} = V_{\text{human reader},i},$$
$$Y_{A,i} = A_{\text{human reader},i}.$$

Valence and Arousal are predicted separately from the layer-wise representations of Reader and Self. Within each training fold, we first standardize each hidden dimension:

$$\widetilde{H} = \text{Standardize}(H).$$

We then apply Principal Component Analysis (PCA) with at most 50 components, subject to the number of training samples and the hidden-state dimensionality:

$$k = \min\left(50, N_{\text{train}}, D\right).$$

When PCA provides an effective dimensionality reduction, the resulting representation

$$Z = \text{PCA}_k(\widetilde{H})$$

is used as input to the probe.

We use Ridge regression to predict each continuous target:

$$\hat{y} = Zw + b,$$

with the objective

$$\min_{w,b} \sum_i (y_i - \hat{y}_i)^2 + \alpha \|w\|_2^2, \qquad \alpha = 1.$$

For validation representations, we apply only the StandardScaler and PCA transformations fitted on the corresponding training fold. To prevent the same text from appearing in both the training and validation subsets, we use 5-fold GroupKFold cross-validation, with normalized stimulus text as the group identifier.

For each layer, task, and affective axis, we evaluate the held-out predictions primarily using

$$R^2$$

and Pearson correlation.

For the Primary V1 decodability summary, we construct layer-wise R^2 profiles separately for Reader and Self and define their peak layers as

$$l_R^* = \arg\max_l R_R^2(l),$$

$$l_S^* = \arg\max_l R_S^2(l).$$

For comparisons across model families, we express these peak locations in terms of relative depth:

$$d_R^* = \frac{l_R^*}{L-1}, \qquad d_S^* = \frac{l_S^*}{L-1}.$$

We additionally report

$$|d_R^* - d_S^*|$$

as an auxiliary measure of the difference in peak positions.

Importantly, proximity between the Reader and Self peak locations alone is not interpreted as evidence of a shared mechanism.

H.5.2 E1-C: AIPsy Affect-Intensity Decodability

For the 48 AIPsy-Affect triplets, the ordered manipulation

$$\text{Neutral} \rightarrow \text{Moderate} \rightarrow \text{Clinical}$$

is encoded as a continuous ordinal target:

$$y = \begin{cases} 0 & \text{Neutral / None,} \\ 1 & \text{Moderate,} \\ 2 & \text{Clinical / Peak.} \end{cases}$$

To ensure that stimuli from the same triplet do not appear across different folds, we use 5-fold GroupKFold cross-validation with triplet identity as the grouping variable. As in the continuous EmoBank probe, the following pipeline is fitted exclusively on the training subset of each fold:

$$\text{StandardScaler} \rightarrow \text{PCA}_{\leq 50} \rightarrow \text{Ridge}(\alpha = 1).$$

For each layer, we record

$$R^2, \qquad \text{Pearson } r, \qquad \text{Spearman } \rho.$$

This analysis serves as a complementary test of whether the hidden states encode not only the binary Clinical–Neutral distinction but also the graded manipulation of affective intensity.

H.5.3 E1-D: EMOTION-CATEGORY DECODABILITY

As a Secondary analysis, we use the eight target emotions associated with the 192 Clinical stimuli to define the multi-class target

$$y_i \in \{\text{rage}, \text{grief}, \text{terror}, \text{loathing}, \text{ecstasy}, \text{admiration}, \text{amazement}, \text{vigilance}\}.$$

For the Reader and Self representations at each layer, we standardize the hidden representations using parameters estimated from the training fold and fit a multi-class Logistic Regression classifier. We evaluate held-out classification performance using

$$\text{macro one-vs-rest ROC-AUC},$$

$$\text{Balanced Accuracy},$$

and

$$\text{Macro F1}.$$

H.6 INTERPRETIVE BOUNDARIES OF E1

We use linear probing to assess information that is linearly readable from hidden representations (Hewitt & Liang, 2019).

Accordingly, positive held-out decodability in E1 directly supports the conclusion that

$$\text{affect-relevant information is linearly readable from the representation.}$$

However, probe performance provides evidence about the decodability of such information; it does not by itself establish that the information at that layer is causally utilized by downstream computation.

Throughout V1, we therefore maintain the interpretive distinction

$$\text{Linear Decodability} \neq \text{Local Causal Utilization}.$$

H.7 E2: SHARED GEOMETRY

Even if the same external target can be decoded from both Reader and Self representations in E1, this does not imply that the two tasks encode the information in the same linear coordinate system. For example, Reader and Self may preserve the same affect-relevant information while their representations correspond only after a simple rotation:

$$H_S \approx H_R Q.$$

E2 therefore compares the representational geometry of Reader and Self using human Reader Valence / Arousal annotations from EmoBank as common external targets. We evaluate this relationship using three complementary analyses:

1. Direct Cross-Decoding
2. Representational Similarity Analysis (RSA) (Kriegeskorte et al., 2008)
3. Orthogonal Procrustes Alignment (Schönemann, 1966)

Under the current Primary design, E2 is conducted on EmoBank.

H.7.1 DIRECT CROSS-DECODING

Let

$$f_R$$

denote a probe trained on Reader representations, and

$$f_S$$

a probe trained on Self representations.

For within-task prediction, we compute

$$R^2_{R\to R}, \qquad R^2_{S\to S}.$$

For cross-task transfer, we compute

$$R^2_{R\to S}, \qquad R^2_{S\to R}.$$

In the Primary Direct Cross-Decoding analysis, the scaler fitted on the source task is applied unchanged to the target-task representations. For example, in Reader-to-Self transfer, the standardization parameters estimated from the Reader training fold are used to transform the held-out Self representations, which are then passed directly to the Reader-trained Ridge probe. This procedure prevents task-specific re-standardization from artificially removing cross-task differences.

We summarize bidirectional cross-decoding as

$$S_{\text{direct}}(l) = \frac{R^2_{R\to S}(l) + R^2_{S\to R}(l)}{2},$$

which serves as an auxiliary measure of direct linear sharing between Reader and Self representations.

H.7.2 Representational Similarity Analysis

We additionally compare the stimulus-to-stimulus relational geometry of the two representation spaces.

For the Reader representations in each held-out fold, we construct a correlation-distance representational dissimilarity matrix (RDM),

$$D_R(i,j) = 1 - \text{Corr}\left(h^R_i, h^R_j\right).$$

Analogously, for Self we construct

$$D_S(i,j) = 1 - \text{Corr}\left(h^S_i, h^S_j\right).$$

We then compute the Spearman correlation between the upper-triangular elements of the two RDMs:

$$\rho_{\text{RSA}}(l) = \text{Spearman}\left(D_R, D_S\right).$$

The held-out RSA values are averaged across cross-validation folds to obtain a layer-wise RSA profile. A high RSA value is consistent with Reader and Self preserving similar relative geometry across stimuli, even when their representations are not expressed along identical coordinate axes.

H.7.3 Orthogonal Procrustes Alignment

We further test whether Reader and Self representations can be aligned through an orthogonal rotation. Within each training fold, the Reader and Self representations are standardized separately, yielding

$$\widetilde{H}_R, \qquad \widetilde{H}_S.$$

We then fit a shared PCA basis to the stacked Reader and Self representations from the training fold. The dimensionality of the shared PCA space is

$$k = \min\left(64, N_{\text{train}} - 1, D\right).$$

Projection into this shared PCA space yields

$$Z_R, \qquad Z_S.$$

Using the training fold, we compute

$$M = Z_S^\top Z_R$$

and perform singular value decomposition,

$$M = U\Sigma V^\top.$$

The orthogonal transformation mapping Self representations toward the Reader representation space is then defined as

$$Q = UV^\top.$$

For a Ridge probe trained on the Reader representations in the shared PCA space, we define the score obtained by applying the probe directly to the unrotated Self representations as

$$R^2_{\mathrm{PCA-direct}},$$

and the score obtained after applying the learned rotation,

$$Z_S Q,$$

as

$$R^2_{\mathrm{aligned}}.$$

We define the alignment gain as

$$\Delta_{\mathrm{align}} = R^2_{\mathrm{aligned}} - R^2_{\mathrm{PCA-direct}}.$$

Both the shared PCA basis and the Procrustes rotation are estimated exclusively from the training fold and then applied to the corresponding held-out fold. E2 therefore distinguishes among direct cross-task transfer, similarity in relational geometry, and linear alignability of Reader and Self representations.

H.8 Phase B: Semantic / Contextual Validity Controls

Even when Phase A yields high decodability, the classifier or regressor may rely primarily on surface-level lexical cues in the stimulus. Phase B therefore tests whether the decodability observed in Phase A can be explained solely by simple surface-level lexical shortcuts, using rule-based controlled perturbations.

We do not use freely generated LLM paraphrases in this phase. Instead, all control stimuli are constructed using prespecified deterministic transformations. This design avoids generating model-specific control stimuli and ensures that all models are evaluated on the same stimulus set.

H.9 Layer Selection for Phase B

Selecting the target layer for Phase B based on an observed peak in Phase A would use the same data for both site selection and subsequent validation. To avoid this form of data-dependent selection, the Primary Phase B analysis does not use the Phase A peak. Instead, we prespecify a relative depth of

$$d = 0.5$$

for all models.

For a model with L Transformer blocks, the target layer is defined as

$$l_{\mathrm{mid}} = \mathrm{round}\left[0.5(L-1)\right].$$

Phase B is therefore not an analysis in which a favorable layer is selected after inspecting the Phase A results. Rather, it serves as an independent validity control conducted at an a priori specified mid-depth representation.

H.10 Phase B Control Stimuli

Using the 192 Clinical–Neutral matched pairs as the starting point, we construct the following stimulus conditions:

1. Original Minimal Pair
2. Paraphrase
3. Word Shuffle
4. Outcome Reversal

H.10.1 ORIGINAL MINIMAL PAIR

We first classify the original Clinical stimuli and their matched Neutral counterparts using the binary target

$$y = \begin{cases} 1 & \text{Clinical / affective,} \\ 0 & \text{Neutral.} \end{cases}$$

To ensure that the Clinical and Neutral stimuli from the same matched pair do not appear in different folds, we use stratified group cross-validation with pair identity as the grouping variable. Within each fold, hidden features are standardized using parameters estimated exclusively from the training subset, and a Logistic Regression classifier is then fitted. The Primary metric is Balanced Accuracy.

This condition serves as a baseline check that the original minimal-pair distinction remains decodable at the prespecified mid-depth layer used in Phase B.

H.10.2 PARAPHRASE INVARIANCE

For each Clinical stimulus, we apply a prespecified set of surface-level lexical substitution rules. The objective is to test whether a classifier trained on the original Clinical–Neutral distinction continues to classify a paraphrased Clinical stimulus as affective when part of its surface wording is altered while broadly preserving the underlying situation.

When no predefined substitution rule is applicable, a fixed reporting clause is appended as a fallback transformation without substantially altering the intended semantic content. Because this fallback differs qualitatively from the intended lexical-substitution manipulation, fallback cases are excluded from the Primary analysis.

H.10.3 OUTCOME REVERSAL

We apply prespecified rule-based polarity reversals to the outcomes described in Clinical stimuli. When a predefined outcome expression is detected, it is replaced with an expression corresponding to the opposite outcome, such as death versus survival, failure versus success, rejection versus acceptance, or guilt versus acquittal.

The Primary question is whether an affective classifier trained on the original Clinical representations assigns a lower affective-class probability to the corresponding outcome-reversed stimuli.

For stimuli to which no predefined reversal rule applies, a fixed resolution clause can be appended as a fallback transformation. Because this fallback introduces substantially more lexical material and cannot be treated as equivalent to a direct outcome reversal, fallback cases are excluded from the Primary analysis.

H.10.4 WORD SHUFFLE

For the Word Shuffle condition, we randomize word order separately within each Clinical and Neutral stimulus while preserving its word set. This manipulation largely preserves lexical identity while disrupting compositional order and syntactic structure.

We perform pair-aware cross-validation on the shuffled Clinical and Neutral representations and compute

$$Acc_{\text{shuffle}}.$$

Let the corresponding accuracy for the Original Minimal Pair condition be

$$Acc_{\text{original}}.$$

We define the auxiliary control quantity

$$\Delta Acc_{\text{shuffle}} = Acc_{\text{shuffle}} - Acc_{\text{original}}.$$

Word Shuffle disrupts not only semantic composition but also syntax and natural sentence structure. Accordingly, any observed decrease in performance is not interpreted as being attributable exclusively to the loss of affective semantics.

H.11 Held-Out Design in Phase B

For the Paraphrase and Outcome Reversal analyses, the same 192 matched pairs are shuffled at the pair level using seed 42 and split into

$$70\% \text{ train} + 30\% \text{ test}.$$

The classifier is trained exclusively on Original Clinical and Original Neutral stimuli from the training set. Thus, neither paraphrased nor outcome-reversed stimuli are observed during training.

For the held-out Paraphrase evaluation, the classifier is evaluated on

$$\{\text{Paraphrased Clinical}, \text{Original Neutral}\}.$$

For Outcome Reversal, we compare, for each held-out Clinical stimulus,

$$P_{\text{aff}}\left(x_{\text{original}}\right)$$

with

$$P_{\text{aff}}\left(x_{\text{reversed}}\right).$$

The Primary reversal quantity is defined as

$$\Delta P_{\text{reversal}} = E\left[P_{\text{aff}}\left(x_{\text{original}}\right)\right] - E\left[P_{\text{aff}}\left(x_{\text{reversed}}\right)\right].$$

A positive value,

$$\Delta P_{\text{reversal}} > 0,$$

indicates that outcome reversal reduces the probability assigned to the affective class.

The Primary held-out evaluation includes only stimuli for which the intended non-fallback transformation is successfully applied. Evaluation over the full set of test pairs, including fallback transformations, is reported separately as a Sensitivity analysis.

H.12 Scope of Evidence in Phase B

Phase B does not establish that the model's representations implement semantic understanding equivalent to that of humans. Nor do the deterministic rule-based controls exhaust all possible lexical confounds.

Rather, Phase B serves as a representational-validity control that tests whether the decodability observed in Phase A changes systematically under selected surface-form perturbations and contextual manipulations.

H.13 Phase C: Causal Intervention

Phase A/B evaluates whether information can be decoded from hidden representations, but decodability alone does not establish that those representations exert a causal influence on output generation.

Building on the general framework of direct interventions on internal activations (Li et al., 2023b; Tak et al., 2025; Banayeeanzade et al., 2026), we construct task- and layer-specific residual interventions.

Phase C directly manipulates hidden representations to test which layers exhibit causal leverage over Reader / Self outputs. It consists of

1. **E3: Shared Causal Map**
2. **E4: Causal Interchangeability**
3. **E6: Task-Specific Causal Specialization**

H.14 Discovery / Confirmation Split

We randomly permute the 192 Clinical–Neutral matched pairs from AIPsy-Affect at the pair level using seed 42 and divide them into

$$96 \text{ Discovery} + 96 \text{ Confirmation}.$$

The two subsets are completely disjoint at the matched-pair level.

The Discovery subset is used to select candidate layers for E4 and task-selective sites for E6. The primary cross-task transfer and task-specialization inferences are then evaluated on the Confirmation subset. This separation prevents the same matched pairs used for site selection from being reused directly as confirmatory evidence.

H.15 E3: Shared Causal Map

E3 asks at which layers intervention on stimulus-related hidden-state differences begins to alter Reader and Self outputs. The production analysis evaluates all Transformer blocks.

H.15.1 Stimulus-Associated Hidden-State Difference

Let

$$T \in \{R, S\}$$

denote the task.

For each Clinical–Neutral matched pair i and layer l, we compute

$$\Delta h_{T,i,l} = h^{C}_{T,i,l} - h^{N}_{T,i,l},$$

where

$$h^{C}_{T,i,l}$$

is the prompt-end representation for the Clinical stimulus and

$$h^{N}_{T,i,l}$$

is the corresponding prompt-end representation for its matched Neutral stimulus.

This difference vector may capture not only the affective manipulation but also pair-specific contextual differences. We therefore do not refer to

$$\Delta h_{T,i,l}$$

as a "pure emotion vector." Instead, we interpret it as an affect-manipulation-associated hidden-state difference.

H.15.2 Same-Task Intervention

For Reader, we intervene on the Neutral representation using the Reader-derived hidden-state difference:

$$h^{N}_{R,i,l} \leftarrow h^{N}_{R,i,l} + \Delta h_{R,i,l}.$$

For Self, we analogously apply

$$h^{N}_{S,i,l} \leftarrow h^{N}_{S,i,l} + \Delta h_{S,i,l}.$$

In E3, the patch weight is fixed to

$$\alpha = 1.$$

The intervention is applied to the prompt-end residual representation at each layer. Following the intervention, we recompute the 729-state VAD candidate distribution and quantify the resulting shifts in expected Valence and Arousal relative to the unpatched Neutral baseline.

For Reader, we define

$$\delta_{R,i,l} = \left(\delta^{V}_{R,i,l}, \delta^{A}_{R,i,l}\right),$$

and for Self,

$$\delta_{S,i,l} = \left(\delta^{V}_{S,i,l}, \delta^{A}_{S,i,l}\right).$$

For example, the Self Valence shift is

$$\delta^{V}_{S,i,l} = E[V]_{\text{patched}} - E[V]_{\text{neutral}}.$$

H.15.3 CAUSAL MAGNITUDE

For matched pair i, we define the joint VA displacement magnitude as

$$C_{T,i}(l) = \|\delta_{T,i,l}\|_2 = \sqrt{(\delta^V_{T,i,l})^2 + (\delta^A_{T,i,l})^2}.$$

The layer-wise profile used for site selection on the Discovery subset is

$$C^D_T(l) = \frac{1}{N_D} \sum_{i \in D} C_{T,i}(l).$$

For the Confirmation subset, we independently retain

$$C^C_T(l) = \frac{1}{N_C} \sum_{i \in C} C_{T,i}(l).$$

Separately, we compute the mean VA shift vector across all matched pairs,

$$\overline{\delta_T}(l) = \frac{1}{N} \sum_i \delta_{T,i,l},$$

and record its magnitude,

$$\left\|\overline{\delta_T}(l)\right\|_2,$$

as an additional layer-wise descriptive measure of causal magnitude.

Thus, we explicitly distinguish the mean of the pairwise displacement magnitudes,

$$\frac{1}{N} \sum_i \|\delta_{T,i,l}\|_2,$$

from the magnitude of the mean shift vector,

$$\left\| \frac{1}{N} \sum_i \delta_{T,i,l} \right\|_2.$$

H.15.4 CAUSAL DIRECTION SIMILARITY

For the mean Reader and Self shift vectors, we compute

$$\cos_l = \frac{\overline{\delta_R}(l)^\top \overline{\delta_S}(l)}{\left\|\overline{\delta_R}(l)\right\|_2 \left\|\overline{\delta_S}(l)\right\|_2}.$$

In addition, for each matched pair, we compute

$$\cos_{i,l} = \cos\left(\delta_{R,i,l}, \delta_{S,i,l}\right),$$

and average this quantity over pairs for which the cosine similarity is well-defined. This analysis evaluates not only whether Reader and Self exhibit large causal magnitudes at similar layers, but also whether interventions at those layers move their VA outputs in similar directions.

H.15.5 READER–SELF CAUSAL-PROFILE COMPARISON

For the Reader and Self causal profiles obtained in E3, we record the peak layer and corresponding relative depth for each task. As a Secondary summary, we additionally compute

$$\rho_C = \mathrm{Spearman}\left(C_R(l), C_S(l)\right),$$

which quantifies the rank similarity between the layer-wise causal-magnitude profiles of Reader and Self.

Accordingly, E3 does not establish that Reader and Self rely on an identical circuit. Rather, it evaluates the extent of Reader–Self overlap in causal sites and the directional similarity of their intervention-induced output shifts.

H.16 E4: Causal Interchangeability

Even if E3 reveals similar causal-site profiles for Reader and Self, this does not imply that the representations formed at those sites are interchangeable across tasks. E4 therefore tests whether a stimulus-associated hidden-state difference formed under Reader can be directly transplanted into the Self computation.

H.16.1 E4 Candidate-Layer Selection

Layers evaluated in E4 are selected without inspecting the Confirmation data. Using the causal-magnitude profiles from the E3 Discovery subset, we identify

$$l_R^* = \arg\max_l C_R^D(l),$$

and

$$l_S^* = \arg\max_l C_S^D(l).$$

To include layers near the identified peaks as well as a late-layer control, we define

$$\mathcal{L}_{E4} = \{l_R^*, l_S^*, l_R^* - 1, l_S^* + 1, \lfloor 0.85L \rfloor\},$$

where indices are clipped to the valid layer range and duplicate entries are removed. Thus, the E4 candidate layers are determined exclusively from the Discovery data.

The Primary confirmatory layer is fixed to

$$l = l_R^*,$$

corresponding to the Reader causal peak identified from the E3 Discovery subset.

H.16.2 Reader-to-Self Matched Transfer

For each Confirmation pair i, we define the Clinical–Neutral difference in the Reader representation as

$$\Delta h_{R,i,l} = h_{R,i,l}^C - h_{R,i,l}^N.$$

We inject this vector into the Self Neutral representation for the same matched pair i:

$$h_{S,i,l}^N \leftarrow h_{S,i,l}^N + \alpha \Delta h_{R,i,l}.$$

The intervention strength is varied over

$$\alpha \in \{0,\ 0.5,\ 1.0,\ 2.0\}.$$

Here,

$$\alpha = 0$$

serves as a zero-intervention sanity condition, whereas

$$\alpha = 1$$

is the reference dose that directly transplants the full Clinical–Neutral hidden-state difference.

H.17 E4 Control Conditions

To test whether Reader-to-Self transfer reflects more than a generic perturbation induced by injecting an arbitrary large vector, we compare the following intervention conditions.

H.17.1 Matched Reader → Self

For each recipient pair i, the Reader-derived difference from the same matched pair,

$$\Delta h_{R,i,l},$$

is injected into the Self Neutral representation. This is the Primary cross-task transfer condition.

H.17.2 RANDOM READER → SELF

Within the same Confirmation subset, each recipient pair i is assigned a Reader-derived difference from a different matched pair. Random donor assignments are constructed as complete derangements with no fixed points:

$$\pi(i) \neq i \qquad \forall i.$$

To avoid dependence on a single random permutation, we generate

$$K = 20$$

independent derangements. For each recipient pair, the intervention effects obtained from the 20 random donors are averaged to define the random-control response.

H.17.3 SELF → SELF

The Self-derived difference from the same matched pair,

$$\Delta h_{S,i,l},$$

is injected into the Self Neutral representation. This condition serves as the same-task reference against which cross-task transfer is compared.

H.17.4 READER → READER

We additionally evaluate a same-task Reader control by injecting

$$\Delta h_{R,i,l}$$

back into the Reader Neutral representation.

H.17.5 SELF → READER

As a reverse cross-task condition, we inject

$$\Delta h_{S,i,l}$$

into the Reader Neutral representation. This condition is retained as a Secondary control for characterizing directional asymmetry in cross-task transfer.

H.18 EMOTION-DIRECTION ALIGNMENT IN E4

Because AIPsy-Affect contains both positive and negative emotions, directly averaging raw signed shifts may lead to cancellation, particularly for Valence. We therefore use the same prespecified emotion-specific directions introduced in the Behavioral Stage,

$$q_{e,d} \in \{-1, +1\}.$$

For example, let

$$\delta_{i,d}^{\text{matched}}$$

denote the matched Reader-to-Self shift for pair i along affective dimension d. We define the direction-aligned shift as

$$\widetilde{\delta}_{i,d}^{\text{matched}} = q_{e_i,d}\delta_{i,d}^{\text{matched}}.$$

The same direction alignment is applied to the Random Reader-to-Self, Self-to-Self, and Reader-to-Reader conditions. Primary E4 effects are defined in terms of these direction-aligned quantities.

H.19 E4 PRIMARY SPECIFICITY

We define the Primary cross-task specificity effect for affective dimension d as

$$\text{Specificity}_d = E\left[\widetilde{\delta}_d^{\text{matched}}\right] - E\left[\widetilde{\delta}_d^{\text{random}}\right].$$

A positive value indicates that a Reader-derived hidden-state difference obtained from the same stimulus pair shifts the Self output more strongly in the prespecified affective direction than does an unmatched Reader-derived difference from a random donor pair.

H.20 E4 Statistical Inference

For each Confirmation pair, we pair the matched intervention effect with the mean random-control effect across the 20 derangements and define the pair-level specificity as

$$\widetilde{\delta}_{i,d}^{\text{matched}} - \widetilde{\delta}_{i,d}^{\text{random}}.$$

For each layer, intervention strength α, and affective dimension, we compute

1. a paired t-test,
2. paired Cohen's d_z,
3. a 95% confidence interval for the mean specificity using 1,000 pair-level bootstrap resamples, and
4. a paired sign-flip permutation test with 10,000 permutations.

Benjamini–Hochberg FDR correction is applied across the multiple layer $\times\alpha$ conditions. The Primary confirmatory condition is fixed to

$$l = l_R^*, \qquad \alpha = 1.$$

The remaining candidate layers and intervention strengths are treated as Secondary analyses for characterizing dose and site dependence.

H.21 E4 Transfer Ratio

As an auxiliary measure of the magnitude of matched Reader-to-Self transfer relative to Self's own same-task causal effect, we compute

$$TR_d = \frac{E\left[\widetilde{\delta}_d^{R\to S}\right]}{E\left[\widetilde{\delta}_d^{S\to S}\right]}.$$

To avoid unstable ratios when the denominator is close to zero, the transfer ratio is left undefined whenever

$$\left|E\left[\widetilde{\delta}_d^{S\to S}\right]\right| < 0.05.$$

H.22 Interpretive Boundaries of E4

Even if matched Reader-to-Self transfer exceeds the random-donor control in E4, we do not interpret

$$\Delta h_R$$

as a context-independent "pure affect code." Rather, Δh_R remains a pair-specific hidden-state difference associated with the Clinical–Neutral stimulus manipulation.

Conversely, failure to observe an E4 effect does not establish that Reader and Self rely on completely independent internal systems. Direct patching may produce off-manifold representations, and the mapping between the two tasks may instead be nonlinear or distributed across multiple representational components. Accordingly, a null E4 result supports only the narrower conclusion that direct cross-task interchangeability was not established under the intervention tested here.

H.23 E6: Task-Specific Causal Specialization

E3 and E4 primarily test the extent to which Reader and Self exhibit shared causal structure. However, even in the presence of such shared structure, the two tasks may exhibit additional task-selective site dependence. E6 therefore tests whether intervention effects at Reader-selective and Self-selective sites differ as a function of task.

H.24 E6 Site Selection

Site selection uses only the E3 Discovery subset. At each layer, we define the task-selectivity contrasts

$$S_R(l) = C_R^D(l) - C_S^D(l)$$

and

$$S_S(l) = C_S^D(l) - C_R^D(l).$$

The Reader-selective site is defined as

$$l_R = \arg\max_l S_R(l),$$

and the Self-selective site as

$$l_S = \arg\max_l S_S(l).$$

We conclude that no distinct task-selective sites are identified if any of the following conditions holds:

$$l_R = l_S,$$

or

$$\max_l S_R(l) \leq 0,$$

or

$$\max_l S_S(l) \leq 0.$$

In such cases, we do not replace the selected site post hoc with an alternative such as a secondary peak. Instead, we record the outcome as "No distinct task-selective sites identified" and do not perform the targeted E6 ablation. Thus, E6 is explicitly designed to permit the possibility that distinct task-selective sites are not identified.

H.25 E6 Targeted Ablation

Only when distinct task-selective sites are identified do we perform a 2×2 factorial intervention on the independent Confirmation pairs (Table 16).

Table 16: The 2×2 targeted-ablation design used in E6.

Task	Ablated Site	Site Type
Reader	Reader-selective layer	ReaderSite
Self	Reader-selective layer	ReaderSite
Reader	Self-selective layer	SelfSite
Self	Self-selective layer	SelfSite

The intervention applies a whole-residual zero ablation to the prompt-end residual representation of each Clinical stimulus:

$$h_{l,t_{\text{end}}} \leftarrow \mathbf{0}.$$

This intervention therefore does not selectively remove an emotion-specific direction. Rather, E6 evaluates task-specific sensitivity to disrupting the entire residual state at a prespecified task-selective site.

H.26 E6 Outcome

For each task, the intact Clinical response before ablation serves as the baseline. We define the Valence shift as

$$\Delta V = E[V]_{\text{ablated}} - E[V]_{\text{intact}},$$

and the Arousal shift as

$$\Delta A = E[A]_{\text{ablated}} - E[A]_{\text{intact}}.$$

The axis-specific impacts are defined as

$$I_V = |\Delta V|,$$

and

$$I_A = |\Delta A|.$$

The Primary joint outcome is

$$I_{VA} = \sqrt{(\Delta V)^2 + (\Delta A)^2}.$$

Thus, the Primary outcome measures the magnitude of the ablation-induced change in joint VA output, irrespective of the direction of that change.

H.27 E6 Statistical Model

The Primary statistical model is a Linear Mixed-Effects Model of the form

$$I_{VA} \sim \text{Task} \times \text{SiteType} + (1|\text{pair}),$$

where Task has the levels

$$\{\text{Reader}, \text{Self}\},$$

and SiteType has the levels

$$\{\text{ReaderSite}, \text{SelfSite}\}.$$

The Primary inferential term is the

$$\text{Task} \times \text{SiteType}$$

interaction, which tests whether the impact of ablating a Reader-selective versus Self-selective site differs between the Reader and Self tasks.

The same model is additionally fitted to the Valence-specific impact I_V and Arousal-specific impact I_A as Secondary analyses.

H.28 E6 Fallback Analysis

If the Linear Mixed-Effects Model cannot be fitted reliably because of numerical issues, we use a prespecified pair-level double-difference analysis as a fallback. For pair i, we define

$$\Delta\Delta_i = [I_i(R, R_s) - I_i(S, R_s)] - [I_i(R, S_s) - I_i(S, S_s)],$$

where

$$R_s$$

denotes the Reader-selective site and

$$S_s$$

denotes the Self-selective site.

Under the fallback analysis, we perform a one-sample t-test of

$$H_0 : E[\Delta\Delta] = 0.$$

This provides a pair-level double-difference test corresponding to the 2×2 Task $\times$ SiteType interaction.

H.29 Interpretive Boundaries of E6

Even if a Task $\times$ SiteType interaction is observed, we do not conclude that Reader and Self rely on two completely independent emotion circuits. The intervention used in E6 applies whole-residual zeroing and therefore disrupts not only affect-related information but also other information encoded in the residual state at the selected site.

Accordingly, the appropriate interpretation is evidence of task-specific causal-site sensitivity or partial causal specialization, rather than fully distinct causal circuits.

Table 17: Primary evidence and Secondary analyses in V1.

Analysis	**Primary Evidence**	**Secondary / Control**
E1 Decodability	EmoBank Reader/Self human-V/A held-out R^2; AIPsy Clinical–Neutral probe	Intensity decoding; emotion-category decoding; peak-distance description
E2 Geometry	Direct cross-decoding; RSA; Procrustes alignment gain	task-specific rescaling variants; layer-wise descriptive profiles
Phase B	Original minimal-pair decoding; held-out non-fallback paraphrase; held-out non-fallback outcome reversal	Word Shuffle; fallback-inclusive sensitivity analyses
E3 Causal Map	Reader / Self layer-wise VA causal magnitude	causal-profile rank correlation; directional cosine similarity
E4 Interchangeability	Discovery Reader-peak layer, $\alpha = 1$, direction-aligned matched-minus-random specificity	other layers / doses; same-task controls; reverse Self-to-Reader; transfer ratio
E6 Specialization	Task $\times$ SiteType interaction on I_{VA}	axis-specific I_V, I_A; double-difference fallback

H.30 Organization of Primary and Secondary Evidence in V1

In V1, we do not treat all descriptive quantities as providing the same level of evidence (Table 17).

I V1 Results

Table 18: V1 E1 (Localization of Representational Decodability Peaks): Peak linear decoding performance (R^2, Pearson r) and corresponding optimal relative depth d^* for Reader Prediction and Self-Report in each model family.

			Valence Peak		**Arousal Peak**		**Reader–Self**
Family	**Variant**	**Task**	Layer (d^*)	R^2 (r)	Layer (d^*)	R^2 (r)	Distance Δd^*
Qwen 2.5 1.5B	Base	Reader	L22 (0.81)	0.378 (0.617)	L24 (0.89)	0.130 (0.369)	V:0.00, A:0.07
		Self	L22 (0.81)	0.391 (0.626)	L26 (0.96)	0.123 (0.356)	
	Instruct	Reader	L15 (0.56)	0.215 (0.467)	L17 (0.63)	0.073 (0.301)	V:0.07, A:0.11
		Self	L13 (0.48)	0.285 (0.536)	L20 (0.74)	0.075 (0.300)	
Llama 3.2 1B	Base	Reader	L14 (0.93)	0.239 (0.490)	L4 (0.27)	0.068 (0.288)	V:0.20, A:0.00
		Self	L11 (0.73)	0.243 (0.494)	L4 (0.27)	0.065 (0.282)	
	Instruct	Reader	L6 (0.40)	0.317 (0.564)	L8 (0.53)	0.113 (0.352)	V:0.13, A:0.00
		Self	L8 (0.53)	0.355 (0.597)	L8 (0.53)	0.106 (0.346)	
Gemma 3 1B	Base	Reader	L16 (0.64)	0.331 (0.576)	L15 (0.60)	0.043 (0.251)	V:0.00, A:0.04
		Self	L16 (0.64)	0.330 (0.576)	L16 (0.64)	0.044 (0.254)	
	Instruct	Reader	L12 (0.48)	0.301 (0.552)	L12 (0.48)	0.094 (0.323)	V:0.08, A:0.36
		Self	L14 (0.56)	0.375 (0.613)	L21 (0.84)	0.088 (0.321)	
OLMo 2 1B	Base	Reader	L14 (0.93)	0.418 (0.647)	L7 (0.47)	0.089 (0.314)	V:0.00, A:0.13
		Self	L14 (0.93)	0.458 (0.677)	L9 (0.60)	0.106 (0.338)	
	Instruct	Reader	L7 (0.47)	0.327 (0.573)	L11 (0.73)	0.085 (0.307)	V:0.00, A:0.27
		Self	L7 (0.47)	0.352 (0.594)	L7 (0.47)	0.086 (0.310)	

Note: Reader and Self decodability peaks occurred at similar relative depths in many conditions, although some model / axis combinations exhibited differences of at least 0.2 in relative depth. Thus, affect-relevant information is decodable from both tasks, while the depth at which it is most strongly linearly decodable does not always coincide.

Table 19: V1 E2 (Shared Representational Geometry): Geometric alignment between Reader and Self representations. We report direct cross-task transfer performance (Direct Transfer R^2), representational similarity (RSA ρ), transfer performance after Procrustes alignment (Aligned R^2), and the resulting Alignment Gain.

Family	Variant	Axis	Direct R→S R^2	Direct S→R R^2	RSA ρ	Aligned R^2	Alignment Gain
Qwen 2.5 1.5B	Base	Valence	−52.885	−20.468	0.900	0.162	0.590
		Arousal	−47.760	−25.065	0.851	0.065	0.435
	Instruct	Valence	−9.094	−8.785	0.810	0.144	1.010
		Arousal	−11.702	−18.340	0.928	−0.001	0.171
Llama 3.2 1B	Base	Valence	−5.998	−4.528	0.850	0.240	0.234
		Arousal	−5.073	−8.614	0.950	0.021	0.158
	Instruct	Valence	−20.771	−14.748	0.802	0.136	1.128
		Arousal	−38.303	−11.039	0.822	−0.033	1.533
Gemma 3 1B	Base	Valence	−2.947	−6.518	0.935	0.161	0.080
		Arousal	−9.795	−5.232	0.928	0.007	0.065
	Instruct	Valence	−16.130	−28.941	0.872	0.241	0.403
		Arousal	−13.957	−9.147	0.756	−0.002	0.281
OLMo 2 1B	Base	Valence	−8.321	−1.437	0.866	0.389	0.707
		Arousal	−9.701	−39.854	0.901	0.086	0.379
	Instruct	Valence	−4.947	−3.487	0.942	0.294	0.190
		Arousal	−57.238	−4.508	0.932	0.062	0.119

Note: Positive RSA was observed between Reader and Self stimulus geometries in all conditions, with many conditions exhibiting high rank-level similarity. In contrast, direct cross-decoding yielded strongly negative R^2 values in many conditions, providing no evidence that the raw coordinate systems are directly interchangeable. Performance generally improved after Procrustes alignment, consistent with the two tasks having representational geometries that are partially alignable through a linear transformation.

Table 20: V1 Phase B (Semantic and Contextual Validity Controls): Probing performance on minimal-pair controls designed to reduce reliance on surface-level lexical and syntactic cues. We report performance on the Original stimuli, Paraphrase controls, Word Shuffle controls and the corresponding Shuffle Drop, and Outcome Reversal controls with the number of non-fallback examples ($N_{\text{nonfallback}}$).

Family	Variant	Task	Original Acc	Paraphrase (N)	Word Shuffle	Shuffle Drop	Reversal Drop (N)
Qwen 2.5 1.5B	Base	Reader	0.891	1.000 ($N = 8$)	0.768	0.122	0.313 ($N = 18$)
		Self	0.904	1.000 ($N = 8$)	0.807	0.096	0.122 ($N = 18$)
	Instruct	Reader	0.885	1.000 ($N = 8$)	0.753	0.133	0.012 ($N = 18$)
		Self	0.859	1.000 ($N = 8$)	0.760	0.099	0.045 ($N = 18$)
Llama 3.2 1B	Base	Reader	0.898	0.750 ($N = 8$)	0.792	0.107	0.019 ($N = 18$)
		Self	0.888	1.000 ($N = 8$)	0.797	0.091	0.013 ($N = 18$)
	Instruct	Reader	0.940	1.000 ($N = 8$)	0.833	0.107	0.037 ($N = 18$)
		Self	0.940	1.000 ($N = 8$)	0.831	0.109	0.044 ($N = 18$)
Gemma 3 1B	Base	Reader	0.875	1.000 ($N = 8$)	0.734	0.141	0.021 ($N = 18$)
		Self	0.867	1.000 ($N = 8$)	0.721	0.146	0.010 ($N = 18$)
	Instruct	Reader	0.914	1.000 ($N = 8$)	0.844	0.070	0.048 ($N = 18$)
		Self	0.922	1.000 ($N = 8$)	0.844	0.078	0.084 ($N = 18$)
OLMo 2 1B	Base	Reader	0.906	1.000 ($N = 8$)	0.779	0.128	0.086 ($N = 18$)
		Self	0.917	1.000 ($N = 8$)	0.797	0.120	0.046 ($N = 18$)
	Instruct	Reader	0.914	1.000 ($N = 8$)	0.812	0.102	0.000 ($N = 18$)
		Self	0.898	1.000 ($N = 8$)	0.815	0.083	0.032 ($N = 18$)

Note: The consistent accuracy drop under Word Shuffle suggests that decodability is difficult to explain solely by a bag-of-words shortcut. Because the Paraphrase and Outcome Reversal analyses contain relatively few non-fallback examples, we interpret them as supplementary evidence for semantic and contextual validity rather than as definitive tests.

Table 21: V1 E3 (Shared Causal Map): Reader and Self output-response profiles under interventions using stimulus-associated hidden-state difference vectors. We report the peak intervention layer, causal-effect magnitude ($\times 10^{-3}$), rank correlation between the layer-wise causal profiles (Spearman ρ_{rank}), and mean directional cosine similarity.

			Reader Peak		**Self Peak**		**Causal-Profile Sharing**	
Family	**Variant**	**Axis**	Layer (d^*)	Mag ($\times 10^{-3}$)	Layer (d^*)	Mag ($\times 10^{-3}$)	Spearman ρ_{rank}	Mean Cosine
Qwen 2.5 1.5B	Base	Valence	L16 (0.59)	2.56	L16 (0.59)	2.46	0.805	0.390
	Instruct	Valence	L1 (0.04)	0.98	L8 (0.30)	0.86	0.799	0.000
Llama 3.2 1B	Base	Valence	L4 (0.27)	0.75	L6 (0.40)	1.19	0.582	0.000
	Instruct	Valence	L6 (0.40)	2.19	L6 (0.40)	1.92	0.882	0.414
Gemma 3 1B	Base	Valence	L2 (0.08)	2.87	L2 (0.08)	3.70	0.802	0.712
	Instruct	Valence	L11 (0.44)	4.89	L11 (0.44)	8.21	0.646	0.266
OLMo 2 1B	Base	Valence	L4 (0.27)	6.14	L4 (0.27)	1.66	0.715	0.210
	Instruct	Valence	L10 (0.67)	0.82	L10 (0.67)	1.17	0.771	0.000

Note: Reader and Self exhibited moderate-to-high positive rank correlations between their layer-wise causal profiles, although the strength of this correspondence varied across models, and some conditions showed directional cosine similarity close to zero. Thus, the two tasks partially share layer-level causal sensitivity, but these results do not establish that they rely on the same causal direction or an identical causal circuit.

Table 22: V1 E4 (Causal Interchangeability): Causal effects of transferring Reader-derived stimulus-associated hidden-state difference vectors into Self computation. We report the matched-transfer effect (Matched Effect), random-transfer control effect (Random Mean), specificity (Specificity $= M - R$), 95% confidence interval, significance (p / FDR q), and Transfer Ratio.

Family	**Variant**	**Axis**	**Matched** ($\times 10^{-3}$)	**Random** ($\times 10^{-3}$)	**Specificity** ($\times 10^{-3}$)	**95% CI** ($\times 10^{-3}$)	p / q	**Transfer Ratio**
Qwen 2.5 1.5B	Base	Valence	0.752	0.311	0.441	[-0.56, 1.42]	0.398 / 1.000	—
		Arousal	-0.109	-0.038	-0.071	[-0.26, 0.13]	0.492 / 0.985	—
	Instruct	Valence	0.635	0.355	0.280	[-0.33, 0.89]	0.357 / 0.762	—
		Arousal	-0.182	-0.156	-0.027	[-0.33, 0.30]	0.867 / 1.000	—
Llama 3.2 1B	Base	Valence	-0.229	-0.014	-0.214	[-0.79, 0.37]	0.459 / 1.000	—
		Arousal	-0.276	-0.126	-0.150	[-0.34, 0.05]	0.132 / 0.353	—
	Instruct	Valence	1.058	1.031	0.027	[-0.94, 0.91]	0.954 / 1.000	—
		Arousal	-0.211	-0.028	-0.183	[-0.52, 0.22]	0.319 / 0.639	—
Gemma 3 1B	Base	Valence	-0.777	-0.107	-0.670	[-5.00, 2.70]	0.717 / 1.000	—
		Arousal	0.099	-0.085	0.184	[-2.06, 2.20]	0.866 / 1.000	—
	Instruct	Valence	1.481	-0.008	1.489	[-0.96, 3.81]	0.224 / 0.896	—
		Arousal	0.459	0.476	-0.017	[-0.97, 0.90]	0.973 / 1.000	—
OLMo 2 1B	Base	Valence	-0.621	-0.401	-0.220	[-0.77, 0.35]	0.453 / 1.000	—
		Arousal	-0.270	-0.219	-0.051	[-0.30, 0.21]	0.727 / 1.000	—
	Instruct	Valence	0.081	0.060	0.022	[-0.57, 0.61]	0.946 / 1.000	—
		Arousal	-0.253	-0.023	-0.230	[-0.52, 0.04]	0.109 / 0.354	—

Note: Matched-minus-random specificity for transferring Reader-derived stimulus-associated hidden-state differences into Self was not significant after FDR correction in any Primary condition. Thus, this analysis did not provide robust evidence for cross-task causal interchangeability of Reader-derived representations.

Table 23: V1 E6 (Task-Specific Causal Specialization): Identification of task-selective causal sites (Selective Depths) for Reader and Self, ablation effects at each site, and the Task × SiteType interaction effect (β_{int}, p, FDR q).

			Reader Site Effect		Self Site Effect		Interaction		
Family	**Variant**	**Selective Sites**	On Reader	On Self	On Reader	On Self	β_{int}	p	FDR q
Qwen 2.5 1.5B	Base	(L1, L17)	0.065	0.047	0.032	0.027	0.0124	0.005	0.010
	Instruct	(L6, L27)	0.033	0.022	0.000	0.000	0.0109	7.04e-07	1.88e-06
Llama 3.2 1B	Base	(L6, L7)	0.005	0.004	0.007	0.008	0.0010	0.035	0.056
	Instruct	(L2, L15)	0.064	0.050	0.000	0.000	0.0136	3.03e-10	1.21e-09
Gemma 3 1B	Base	(L0, L18)	0.056	0.052	0.006	0.006	0.0041	0.236	0.236
	Instruct	(L4, L11)	0.067	0.081	0.059	0.061	-0.0121	0.057	0.067
OLMo 2 1B	Base	(L4, L13)	0.066	0.008	0.002	0.003	0.0583	1.84e-135	1.48e-134
	Instruct	(L5, L10)	0.012	0.011	0.004	0.005	0.0018	0.058	0.067

Note: The Task × SiteType interaction was nominally significant for Qwen Base / Instruct, Llama Base / Instruct, and OLMo Base, whereas clear support was not observed for Gemma Base / Instruct or OLMo Instruct. Thus, task-specific causal specialization was observed in some models, but was not a consistent property across all model families.

J V1 Discussion

The V1 results suggest that the Reader–Self coupling observed at the Behavioral Stage extends beyond correlations in final outputs: affect-relevant information is internally represented in both tasks. At the same time, the results do not support a simple account in which the two tasks rely on identical representations or identical causal utilization.

First, in E1, human Reader Valence / Arousal could be decoded from both Reader and Self hidden states on held-out data across both Base and Instruct variants of the four Primary model families. For many model / axis combinations, the peak depths for Reader and Self were also close.

This pattern is more consistent with affect-relevant information being retained in the internal representations of both tasks than with an account in which Reader and Self perform entirely independent information processing. However, some conditions showed differences in peak relative depth exceeding 0.2. Thus,

$$\text{Shared Affect-Relevant Information} \neq \text{Identical Readout Location.}$$

E2 further showed positive RSA between Reader and Self stimulus geometries across all conditions, whereas direct cross-decoding was strongly negative in many conditions. Performance improved after Procrustes alignment, suggesting that Reader and Self representations are not directly interchangeable in their raw coordinate systems but instead exhibit relational geometries that can be partially aligned through a linear transformation. This distinction can be summarized as

$$\text{Similar Relational Geometry} \neq \text{Same Coordinate System.}$$

The Phase B semantic controls showed a consistent decline in performance under Word Shuffle. The V1 decodability results are therefore difficult to explain solely in terms of simple bag-of-words lexical shortcuts. By contrast, the numbers of nonfallback samples available for the Paraphrase and Outcome Reversal controls were small; we therefore interpret these analyses more narrowly as supplementary evidence for semantic validity.

In Phase C, E3 revealed moderate-to-high positive rank correlations between the layer-wise causal profiles of Reader and Self. In many models, their causal peaks also occurred at the same or nearby layers. Thus, Reader and Self appear not only to contain decodable affect-relevant information but also to partially share their layer-level sensitivity to hidden-state perturbations.

However, directional cosine similarity varied substantially across models, including conditions in which it was close to zero. Therefore,

$$\text{Similar Causal Sensitivity Profile} \neq \text{Same Causal Direction.}$$

This distinction became more pronounced in E4. Matched interventions that transferred Reader-derived stimulus-associated hidden-state differences into Self did not show robust specificity relative to the random-transfer control after FDR correction. Accordingly,

$$\text{Reader Representation} \not\Rightarrow \text{Causally Interchangeable Self Representation.}$$

This negative result is important for the interpretation of the broader study. Behavioral coupling, linear decodability, similarity in representational geometry, and similarity in layer-wise causal profiles are not sufficient to establish that Reader and Self directly share the same causal code.

Finally, in E6, the Task $\times$ SiteType interaction was supported in some models, but the pattern did not generalize consistently across all model families. The results are therefore compatible with partial causal specialization between Reader and Self computations, but we do not interpret them as evidence for a family-general architecture consisting of two distinct affective circuits.

Taken together, V1 supports a structured account in which Reader and Self partially share affect-relevant information and layer-level causal sensitivity, while their representations are not directly interchangeable and their causal organization exhibits task-specific differences.

K From V1 to V2

V1 showed that affect-relevant information is present in both Reader and Self, with partial sharing in representational geometry and layer-wise causal sensitivity.

At the same time, direct cross-decoding was weak, and cross-task causal interchangeability was not supported. Moreover, peak depth, causal magnitude, intervention direction, and task-specific site sensitivity varied across models and between Base and Instruct variants.

These findings motivate the next question: which aspects of the representation–report relationship differ between Base and Instruct models?

Importantly, because V1 identified affect-relevant information in both Base and Instruct models, post-training-associated behavioral differences cannot necessarily be explained by simple information erasure. At least four forms of reorganization can be distinguished:

1. transformation of representational geometry,
2. reorganization of Reader–Self sharing,
3. relocation of causal leverage, and
4. changes in the readout from internal representations to the output distribution.

Accordingly, V2 shifts the primary comparison from

$$\text{Reader} \leftrightarrow \text{Self}$$

to

$$\text{Base} \leftrightarrow \text{Instruct.}$$

Using the Matched-Plain condition as the Primary comparison to control for prompt-format differences, V2 examines post-training-associated reorganization of the relationship between affect-relevant internal representations and model outputs.

L V2 Overview

The outline of V2 stage is shown in Fig. 4.

L.1 Objective and Positioning

The central Research Question of V2 is: "How are the representational geometry of affect-relevant information, Reader–Self sharing, the layer-wise distribution of causal leverage, and readout into model outputs reorganized in association with the transition from Base to Instruct models?"

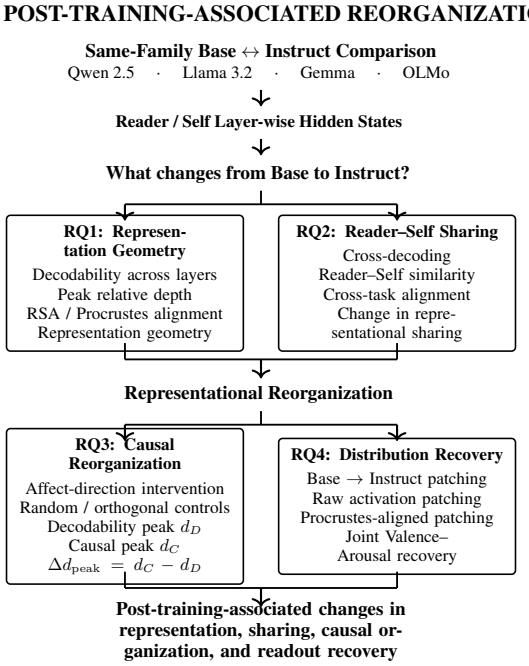


Figure 4: Overview of V2. Within each model family, Base and Instruct variants are compared to characterize post-training-associated reorganization. RQ1 examines changes in affective representational geometry, and RQ2 evaluates changes in Reader–Self representational sharing. RQ3 compares layer-wise decodability with causal effects measured by affect-direction interventions and control directions. RQ4 tests whether Base-like joint valence–arousal distributions can be recovered in the Instruct model through activation patching.

The primary comparison is between pretrained and post-trained models within the same model family:

$$\text{Base} \leftrightarrow \text{Instruct} \quad \text{within the same model family.}$$

Throughout V2, we strictly distinguish between across-model observational comparisons of open-weight models and within-model causal interventions performed on a fixed model.

L.2 Evidence Structure in V2

V2 is organized around four Research Questions:

$$\begin{aligned}&\text{RQ1: Representation Geometry}\\ &\quad\downarrow\\ &\text{RQ2: Reader–Self Sharing}\\ &\quad\downarrow\\ &\text{RQ3: Decodability–Causal Relocation}\\ &\quad\downarrow\\ &\text{RQ4: Cross-Model Distribution Recovery.}\end{aligned}$$

Table 24 summarizes the role of each RQ.

Rather than reducing post-training-associated change to a binary account in which information is simply "preserved" or "lost," V2 decomposes it into

$$\boxed{\text{Representation} + \text{Geometry} + \text{Sharing} + \text{Causal Localization} + \text{Recoverability}}.$$

Table 24: The four Research Questions comprising V2.

RQ	Analysis	Question
RQ1	Representation Geometry	How do the decodability of affect-relevant information and Base–Instruct representational geometry differ?
RQ2	Reader–Self Sharing	How does cross-decoding from Reader to Self and from Self to Reader change between Base and Instruct?
RQ3	Decodability–Causal Relocation	Do the layers at which information is decodable coincide with the layers at which interventions alter the output, and how does this relationship change between Base and Instruct?
RQ4	Distribution Recovery	To what extent can transplanting a Base hidden state into Instruct shift the Instruct VA output distribution back toward the Base distribution?

This decomposition allows us to characterize post-training-associated reorganization at multiple levels of the representation–output relationship.

L.3 Data, External Labels, and Splits

V2 uses the 1,000 stimuli from the EmoBank official test split defined in the preceding section. The external affective targets in V2 are the human Reader annotations from EmoBank:

$$Y_V = V_{\text{human reader}},$$

and

$$Y_A = A_{\text{human reader}}.$$

The same human Reader V/A annotations are used as external targets when probing both Reader and Self representations. Thus, decodability in V2 measures the extent to which the same external human-affect variable is linearly represented in the hidden states of Reader or Self. The model's own Self-Report is not used as the probe target.

L.3.1 Train–Test Split for RQ1/RQ2

For RQ1 and RQ2, the 1,000 stimuli are randomly permuted using seed 42 and divided into

$$70\% \text{ training} + 30\% \text{ held-out test.}$$

Because the current EmoBank evaluation set does not provide matched-pair identifiers, this procedure uses a seeded split at the stimulus-row level. Accordingly,

$$N_{\text{train}} = 700, \qquad N_{\text{test}} = 300.$$

The same split is shared across all model families, Base / Instruct conditions, and Reader / Self tasks. This ensures that comparisons across models and prompt conditions are evaluated on identical stimulus subsets.

L.4 Prompt-Format Control and Internal Representation Measurement

Following the prompt-format controls defined in §C.7—Base Plain, Instruct Matched-Plain, and Instruct Native-Chat—the Primary V2 comparison is the prompt-format-matched contrast

$$\text{Base Plain} \quad \text{vs.} \quad \text{Instruct Matched-Plain}.$$

Comparisons involving Instruct Native-Chat are treated as Secondary analyses. In particular, the contrast

$$\text{Instruct Native-Chat} - \text{Instruct Matched-Plain}$$

is analyzed separately as a prompt-format effect.

Following the common conventions in §C, we extract the prompt-end residual representation

$$h_{M,T,i,l}, \qquad T \in \{R, S\}, \qquad M \in \{\text{Base}, \text{Instruct}\},$$

at each Transformer block l. Internal representations are measured for both Reader and Self, and relative depth

$$d_l = \frac{l}{L-1}$$

is used consistently for cross-model comparisons.

L.5 RQ1: Representation Geometry

RQ1 asks how the location and geometry of affect-relevant representations change in association with the transition from Base to Instruct. We characterize these changes using layer-wise affect decodability, Representational Similarity Analysis (RSA) (Kriegeskorte et al., 2008), and held-out Orthogonal Procrustes disparity.

L.6 RQ1-A: Layer-Wise Affect Decodability

For each Reader and Self layer representation, we predict human Reader Valence and Arousal. In V1 Phase A, the probing pipeline was

$$\text{StandardScaler} \rightarrow \text{PCA} \rightarrow \text{Ridge}.$$

In contrast, V2 RQ1/RQ2 uses

$$\text{raw hidden state} \rightarrow \text{Ridge}$$

to avoid removing post-training-associated changes in the raw representation through preprocessing. Neither StandardScaler nor PCA is applied before the Primary Ridge probe.

At each layer, we fit

$$\hat{y} = H_l w + b$$

using

$$\text{Ridge}(\alpha = 1)$$

on the 700 training stimuli and evaluate

$$R^2(l)$$

on the 300 held-out stimuli.

This procedure yields layer-wise Valence and Arousal decodability profiles for Base Reader, Base Self, Instruct Native-Chat Reader / Self, and Instruct Matched-Plain Reader / Self.

L.7 Decodability Peak

For task T, let

$$d^*_{D,B,T}$$

denote the relative depth of the decodability peak in the Base profile. For Matched-Plain Instruct, we analogously define

$$d^*_{D,I,T}.$$

We characterize the post-training-associated shift in peak decodability depth as

$$\Delta d^*_{D,T} = d^*_{D,I,T} - d^*_{D,B,T}.$$

A positive value indicates that the maximum decodability in Instruct occurs at a later relative depth than in Base, whereas a negative value indicates a shift toward earlier layers. The corresponding peak shift is also computed for the Native-Chat condition as a Secondary analysis.

L.8 RQ1-B: Representational Similarity Analysis

Layer-wise decodability profiles alone do not reveal how the same stimuli are arranged relative to one another in the hidden spaces of Base and Instruct. We therefore compute pairwise cosine distances from the held-out test representations separately for each task.

For Base, we define

$$D_B(i,j) = 1 - \mathrm{CosSim}\left(h_{B,i}, h_{B,j}\right),$$

and for Instruct,

$$D_I(i,j) = 1 - \mathrm{CosSim}\left(h_{I,i}, h_{I,j}\right).$$

We then compute the Spearman correlation between the upper-triangular entries of the two Representational Dissimilarity Matrices:

$$\rho_{\mathrm{RSA}} = \mathrm{Spearman}\left(D_B, D_I\right).$$

The Primary comparison uses RSA between

$$\text{Base Plain} \leftrightarrow \text{Instruct Matched-Plain}.$$

As Secondary comparisons, we additionally compute RSA for

$$\text{Base Plain} \leftrightarrow \text{Instruct Native-Chat},$$

and

$$\text{Instruct Matched-Plain} \leftrightarrow \text{Instruct Native-Chat}.$$

A high RSA value indicates that the relative geometry among stimuli is similar across the two representational spaces. Importantly, RSA captures similarity in relational structure and does not require coordinate-wise identity between the representations.

L.9 RQ1-C: Held-Out Orthogonal Procrustes Disparity

Even when Base and Instruct preserve similar information, post-training may be associated with a rotation or deformation of the representational coordinate system. We therefore estimate an orthogonal alignment between Base and Instruct using only the training split and evaluate the remaining representational mismatch after alignment on the held-out test split.

L.9.1 Centering

For a given task and layer, let

$$\mu_B, \qquad \mu_I$$

denote the mean Base and Instruct representations, respectively, estimated from the training split. We center the two representation matrices separately:

$$X_B = H_B - \mu_B,$$

and

$$X_I = H_I - \mu_I.$$

L.9.2 Common PCA

When the hidden dimension exceeds 64, we concatenate the centered Base and Instruct training representations,

$$X_{\mathrm{joint}} = \begin{bmatrix} X_B \\ X_I \end{bmatrix},$$

and fit a common PCA basis to the resulting matrix.

The retained dimensionality is

$$k = \min\left(64, N_{\mathrm{train}} - 1, D\right).$$

Let V_k denote the resulting shared PCA basis. We project the Base and Instruct representations into this common subspace as

$$Z_B = X_B V_k,$$

and

$$Z_I = X_I V_k.$$

L.10 Interpretive Scope of RQ1

RQ1 measures Decodability and Representation Geometry separately. For example, if Instruct preserves a similar level of decodability while exhibiting lower RSA or greater Procrustes disparity, this pattern would be consistent with a reorganization of representational geometry rather than a simple loss of affect-relevant information. Conversely, similarity in representational geometry does not imply that the corresponding representations are utilized in the same way to produce the final output.

L.11 RQ2: Reader–Self Representation Sharing

RQ2 examines how the Reader–Self relationship investigated in V1 changes in association with the transition from Base to Instruct. The central question is: "How does the affect-relevant linear geometry shared between Reader and Self change between Base and Instruct?"

L.12 Bidirectional Cross-Decoding

For each layer and affective axis, we train a Ridge probe to predict the human Reader affect target from Reader representations. We then apply the Reader-trained probe directly to the held-out Self representations to obtain

$$R^2_{R\to S}(l).$$

Conversely, we train the probe on Self representations and apply it directly to the held-out Reader representations to obtain

$$R^2_{S\to R}(l).$$

During cross-decoding, the source and target representations are not separately restandardized on a task-specific basis. Thus, cross-decoding performance reflects the extent to which the Reader and Self hidden-space coordinate systems are directly compatible without an additional task-specific transformation.

L.13 Reader–Self Sharing Score

We define the bidirectional cross-decoding average as

$$S(l) = \frac{R^2_{R\to S}(l) + R^2_{S\to R}(l)}{2}.$$

For Base, we compute

$$S_B(l),$$

and for Matched-Plain Instruct,

$$S_{I,\mathrm{MP}}(l).$$

The Primary post-training-associated change in Reader–Self sharing is defined as

$$\Delta S_{\mathrm{MP}}(l) = S_{I,\mathrm{MP}}(l) - S_B(l).$$

As a summary across layers, we use

$$\overline{\Delta S}_{\mathrm{MP}} = \frac{1}{L}\sum_l \Delta S_{\mathrm{MP}}(l).$$

L.14 Native and Format Controls

For Native-Chat Instruct, we similarly compute

$$S_{I,\mathrm{Native}}(l),$$

and define

$$\Delta S_{\mathrm{Native}}(l) = S_{I,\mathrm{Native}}(l) - S_B(l)$$

as a Secondary comparison.

To isolate the prompt-format effect, we additionally compute

$$\Delta S_{\text{format}}(l) = S_{I,\text{Native}}(l) - S_{I,\text{MP}}(l).$$

This decomposition separates the Base–Instruct difference under matched prompt formatting from the difference associated with the plain–chat prompt format.

L.15 Directional Asymmetry

As a Secondary measure, we characterize whether the Base–Instruct changes in Reader-to-Self and Self-to-Reader cross-decoding are asymmetric. We compute

$$\Delta\Delta_{\text{cross}} = \left[R^2_{R\to S,I} - R^2_{R\to S,B}\right] - \left[R^2_{S\to R,I} - R^2_{S\to R,B}\right].$$

This quantity is not used as the Primary sharing score; rather, it provides an auxiliary characterization of directional asymmetry in cross-task transfer.

L.16 Interpretive Scope of RQ2

If

$$\Delta S_{\text{MP}} > 0,$$

Reader–Self linear transferability is greater for Instruct than for Base under the Matched-Plain condition. Conversely, if

$$\Delta S_{\text{MP}} < 0,$$

the result is consistent with reduced linear coordinate sharing between Reader and Self, even if the same external affective information remains decodable in both tasks.

Cross-decoding, however, remains an observational measure of representational sharing. Accordingly,

$$\text{Cross-decoding} \neq \text{Causal Utilization}.$$

RQ3 therefore directly tests which layers exert causal leverage over the output.

L.17 RQ3: Decodability–Causal Relocation

RQ3 asks whether the layers at which affect-relevant information is most decodable coincide with the layers at which interventions on that information most strongly alter the output. We further examine how the relationship between decodability and causal leverage changes between Base and Instruct. RQ3 analyzes both Reader and Self under Base Plain, Instruct Matched-Plain, and Instruct Native-Chat conditions.

L.18 Cross-Fitted Affect Directions

Rather than using the fixed 70/30 split employed in RQ1/RQ2, RQ3 uses all stimuli through 5-fold cross-fitting. Within each fold, Ridge regression models are fitted using only the training stimuli to predict human Reader Valence and Arousal from the hidden state at layer l.

For Valence, we fit

$$Y_V = H_l w_V + b_V + \epsilon,$$

and for Arousal,

$$Y_A = H_l w_A + b_A + \epsilon.$$

The Ridge regularization parameter in RQ3 is fixed to

$$\alpha_{\text{Ridge}} = 10.$$

We normalize each coefficient vector to unit norm and define the affect-related intervention directions as

$$d_V = \frac{w_V}{\|w_V\|_2},$$

and

$$d_A = \frac{w_A}{\|w_A\|_2}.$$

Each direction is estimated exclusively from the training portion of a fold and is used for intervention only on the held-out stimuli from that fold. Thus, the intervention direction applied to a given stimulus is never fitted using that stimulus's own affect label.

L.19 Intervention Dose Normalization

For the training hidden states, we compute their projections onto the target direction as

$$z_V = H_{\text{train}} d_V.$$

We then define

$$\sigma_V = \text{SD}(z_V).$$

Analogously, for Arousal,

$$\sigma_A = \text{SD}\left(H_{\text{train}} d_A\right).$$

The intervention is applied as

$$h' = h + \alpha \sigma_d d.$$

The Primary intervention dose is

$$\alpha = 1.$$

Thus, the intervention magnitude is normalized not by the raw hidden-coordinate scale, but by approximately one standard deviation of activation along the corresponding direction in the training representations.

L.20 Random and Orthogonal Controls

To control for the possibility that arbitrary hidden-state perturbations produce larger output changes at deeper layers, we construct two control directions in addition to the affect-related direction:

1. a random direction, and
2. a direction orthogonal to the affect-related direction.

The control directions are normalized to have the same Euclidean norm as the target direction. For Valence interventions, the affect-related direction d_V, random direction $d_{V,\text{rand}}$, and orthogonal direction $d_{V,\perp}$ are all applied using the same σ_V and

$$\alpha = 1.$$

The analogous procedure is used for Arousal.

Thus, differences between the affect-related intervention and its controls are less readily attributable to differences in perturbation magnitude alone.

L.21 Axis-Specific Raw Causal Displacement

Following intervention along the Valence direction, let

$$E[V]_{\text{affect}}$$

denote the expected Valence under the affect-direction intervention, and let

$$E[V]_{\text{clean}}$$

denote the corresponding clean baseline. We define the raw Valence causal displacement as

$$C_{V,\text{raw}} = \left|E[V]_{\text{affect}} - E[V]_{\text{clean}}\right|.$$

For the random and orthogonal controls, we analogously define

$$C_{V,\text{rand}}, \qquad C_{V,\perp}.$$

For Arousal, the corresponding raw causal displacement is

$$C_{A,\text{raw}} = \left|E[A]_{\text{affect}} - E[A]_{\text{clean}}\right|.$$

Thus, the raw causal displacement is defined as a non-negative absolute effect.

L.22 Net Causal Effect

The Primary causal metric in RQ3 is not the raw affect-direction effect itself. Instead, we subtract the generic sensitivity induced by a random perturbation of the same normalized dose:

$$C_{V,\mathrm{net},\mathrm{rand}} = C_{V,\mathrm{raw}} - C_{V,\mathrm{rand}},$$

and

$$C_{A,\mathrm{net},\mathrm{rand}} = C_{A,\mathrm{raw}} - C_{A,\mathrm{rand}}.$$

These random-control-adjusted quantities constitute the Primary causal metrics.

As a Secondary robustness measure, we analogously subtract the orthogonal-control effect:

$$C_{V,\mathrm{net},\perp} = C_{V,\mathrm{raw}} - C_{V,\perp},$$

and

$$C_{A,\mathrm{net},\perp} = C_{A,\mathrm{raw}} - C_{A,\perp}.$$

We additionally record the effect of replacing the entire prompt-end residual state with zero as a non-specific Secondary control.

Importantly,

$$C_{\mathrm{net},\mathrm{rand}}$$

is a signed quantity and may take negative values. A negative value indicates that intervention along the affect-related direction produces a smaller displacement than a random perturbation at the same normalized dose. For statistical analyses, the signed net effect is retained as observed and is not clipped at zero.

L.23 Layer-Wise Causal Profile

Through cross-fitting, each stimulus is evaluated exactly once as a held-out sample under the intervention procedure. For each layer, we average the sample-level net causal effects to obtain the layer-wise profiles

$$C_{V,\mathrm{net}}(l)$$

and

$$C_{A,\mathrm{net}}(l).$$

The Primary causal profile is based on

$$C_{\mathrm{net},\mathrm{rand}},$$

that is, the affect-direction effect after subtracting the matched-dose random-perturbation effect.

L.24 Decodability–Causal Dissociation

For a given condition, task, and affective axis, we denote the held-out decodability profile obtained in RQ1/RQ2 as

$$D(l) = R^2(l).$$

The corresponding Primary net causal profile obtained in RQ3 is denoted as

$$C(l) = C_{\mathrm{net},\mathrm{rand}}(l).$$

L.24.1 Positive Decodability Peak

For the decodability profile, if

$$\max_l D(l) \le 0,$$

we define

$$d_D^* = \mathrm{NaN}.$$

Thus, when the held-out R^2 is non-positive at every layer, we do not designate the "least poor" layer as a decodability peak.

L.24.2 Positive Causal Peak

Similarly, if

$$\max_l C(l) \leq 0,$$

we define

$$d_C^* = \mathrm{NaN}.$$

This criterion prevents us from describing a causal peak when the affect-specific perturbation fails to exceed the random-control effect at every layer.

L.24.3 Peak Dissociation

We define the peak dissociation only when both peaks are well-defined:

$$\Delta d^* = d_C^* - d_D^*.$$

A positive value indicates

$$d_C^* > d_D^*,$$

meaning that the layer of maximal causal leverage occurs later in the network than the layer of maximal decodability.

L.25 Center-of-Mass Dissociation

Because a single argmax may not adequately capture shifts in the overall layer profile, we additionally compute a center of mass using the positive profile mass. For decodability, we define

$$\bar{d}_D = \frac{\sum_l d_l \max(D(l), 0)}{\sum_l \max(D(l), 0)}.$$

If no positive decodability mass is present, we set

$$\bar{d}_D = \mathrm{NaN}.$$

For the causal profile, we analogously define

$$\bar{d}_C = \frac{\sum_l d_l \max(C(l), 0)}{\sum_l \max(C(l), 0)}.$$

If no positive net causal mass is present, we set

$$\bar{d}_C = \mathrm{NaN}.$$

When both quantities are defined, we use

$$\Delta \bar{d} = \bar{d}_C - \bar{d}_D$$

as a profile-level measure of decodability–causal dissociation.

L.26 Post-Training-Associated Change in Dissociation

For Base task T, we compute

$$\Delta d_{B,T}^* = d_{C,B,T}^* - d_{D,B,T}^*.$$

For Matched-Plain Instruct, we similarly define

$$\Delta d_{I,T}^* = d_{C,I,T}^* - d_{D,I,T}^*.$$

Their difference,

$$\Gamma_T^{\mathrm{peak}} = \Delta d_{I,T}^* - \Delta d_{B,T}^*,$$

quantifies the post-training-associated change in decodability–causal peak dissociation.

For the center-of-mass measure, we analogously compute

$$\Gamma_T^{\mathrm{CoM}} = \Delta \bar{d}_{I,T} - \Delta \bar{d}_{B,T}.$$

The Primary comparison uses the Matched-Plain condition, whereas the Native-Chat comparison is treated as Secondary.

L.27 Sample-Level Mixed-Effects Analysis

In addition to the peak- and center-of-mass-based summaries, we conduct a sample-level analysis using the full layer-wise causal profile. The Primary outcome is

$$C_{\text{net,rand}}.$$

Valence and Arousal are analyzed separately using the following Linear Mixed-Effects Model:

$$C_{\text{net,rand}} \sim C(\text{family}) + C(\text{alignment}) \times C(\text{task}) \times \text{relative depth} + (1|\text{stimulus}).$$

Here,

$$\text{alignment} \in \{\text{Base}, \text{Instruct}\},$$

and

$$\text{task} \in \{\text{Reader}, \text{Self}\}.$$

Because EmoBank does not contain a matched-pair structure, the random-intercept grouping variable corresponds to individual stimulus identity. The Primary LMM includes only Base Plain and Instruct Matched-Plain conditions. The same model specification using Instruct Native-Chat in place of Instruct Matched-Plain is treated as a Secondary analysis.

L.28 Prespecified RQ3 Interaction Terms

The Primary confirmatory terms are

$$\text{Alignment} \times \text{RelativeDepth},$$

$$\text{Alignment} \times \text{Task},$$

and

$$\text{Alignment} \times \text{Task} \times \text{RelativeDepth}.$$

We apply Benjamini–Hochberg FDR correction to the p-values of these prespecified interaction terms across Valence and Arousal. This analysis evaluates, at the sample level, whether the shape of the causal-leverage profile differs between Base and Instruct, complementing the peak-based summaries.

L.29 Interpretive Boundaries of RQ3

Because RQ3 directly intervenes on hidden states, it provides causal evidence within each fixed model. However,

$$\Gamma_T^{\text{peak}}$$

and the alignment interactions in the LMM compare distinct Base and Instruct model artifacts. We therefore do not interpret these results as showing that post-training caused the causal peak to move. Rather, they characterize differences in causal profiles between Base and Instruct that are associated with post-training.

L.30 RQ4: Cross-Model Distribution Recovery

RQ1–RQ3 compare Base and Instruct in terms of representational geometry, Reader–Self sharing, and causal localization. These differences alone, however, do not establish whether representations naturally formed in Base can recover a Base-like output distribution when directly introduced into the Instruct computation. RQ4 therefore asks: "Can Base hidden representations recover a Base-like VA output distribution when inserted into the Instruct model?"

To address this question, we perform cross-model activation replacement, in which the hidden representation naturally formed by the Base model for a given stimulus is transplanted into the corresponding Instruct model from the same model family.

L.31 RQ4 Train–Evaluation Split

We split the 1,000 EmoBank stimuli using seed 42 into

$$70\% \text{ train} + 30\% \text{ independent evaluation.}$$

Because the current dataset does not contain a matched-pair structure, we use a stimulus-level permutation split with

$$N_{\text{train}} = 700, \qquad N_{\text{eval}} = 300.$$

Within each model family, the same split is used for Reader and Self. The training subset is used exclusively to estimate the Procrustes alignment, whereas recovery is evaluated on the independent set of 300 stimuli.

L.32 Target and Clean Distributions

For stimulus i and task T, we denote the 81-state VA distribution produced by the Base Plain model as

$$P_{B,i}^{T},$$

which serves as the target distribution for recovery.

The clean Instruct distribution depends on the prompt condition. For the Native-Chat condition, it is denoted by

$$P_{I,\text{Native},i}^{T},$$

whereas for the Matched-Plain condition, it is denoted by

$$P_{I,\text{MP},i}^{T}.$$

Thus, the Primary Matched-Plain recovery analysis quantifies the extent to which Base activation patching reduces the distributional discrepancy

$$P_{I,\text{MP}} \to P_B.$$

L.33 Joint VA Earth Mover's Distance

As the Primary distance between output distributions, we use the Earth Mover's Distance over the joint 81-state VA distribution. Each candidate state is represented as

$$c = (v, a), \qquad v, a \in \{1, \ldots, 9\},$$

and the ground cost between two states is defined by their Euclidean distance in VA space:

$$g(c, c') = \sqrt{(v - v')^2 + (a - a')^2}.$$

For two joint distributions,

$$P, \qquad Q,$$

we denote the optimal transport cost under this ground metric by

$$EMD_{VA}(P, Q).$$

This two-dimensional joint EMD serves as the Primary distributional distance in RQ4 (Rubner et al., 2000).

We additionally retain the one-dimensional Wasserstein distance between the Valence marginals, the one-dimensional Wasserstein distance between the Arousal marginals, and the Jensen–Shannon divergence as Secondary distributional diagnostics. The Recovery metric in RQ4, however, is defined using the joint EMD_{VA}.

L.34 Condition A: Direct Native Patching

For each evaluation stimulus, let

$$h_{B,i,l}$$

denote the prompt-end post-MLP residual state at layer l of the Base Plain model. We directly replace the residual state at the corresponding layer and semantic position in the Instruct Native-Chat model with this unmodified Base activation:

$$h_{I,i,l}^{\text{Native}} \leftarrow h_{B,i,l}.$$

This condition evaluates raw cross-model direct interchangeability. However, if the representational coordinate system differs between Base and Instruct, the transplanted Base activation may occupy an unnatural location on the Instruct representational manifold. We therefore treat Direct Native patching as a Secondary condition.

L.35 Primary Condition: Matched-Plain Direct Patching

For the Primary condition, we use a plain prompt for Instruct to control for prompt-format differences. Let

$$h_{B,i,l}$$

denote the Base Plain activation. We directly replace the corresponding activation at the same layer in Instruct Matched-Plain:

$$h_{I,i,l}^{\text{MP}} \leftarrow h_{B,i,l}.$$

Because both Base and Instruct use the same plain prompt format in this condition, Matched-Plain direct recovery serves as the Primary condition for RQ4.

L.36 Mechanistic Control: Procrustes-Aligned Patching

Failure of direct patching does not necessarily imply that the relevant information is absent from the Base representation. If the representational coordinate system has changed in association with post-training, a raw Base activation may be off-manifold when inserted directly into the Instruct model. We therefore introduce a coordinate-remapping control for the Native-Chat condition by learning a linear alignment from Base representations to Instruct representations using only the training split.

L.36.1 Train-Only Centering and Common PCA

Let

$$H_B, \qquad H_I$$

denote the training activations at layer l from Base Plain and Instruct Native-Chat, respectively. We center each representation using its corresponding training mean,

$$\mu_B, \qquad \mu_I,$$

such that

$$X_B = H_B - \mu_B,$$

and

$$X_I = H_I - \mu_I.$$

We then construct the joint training matrix

$$X_{\text{joint}} = \begin{bmatrix} X_B \\ X_I \end{bmatrix}$$

and estimate a common PCA basis V_k. The subspace dimension is

$$k = \min\left(64, N_{\text{train}} - 1, D\right).$$

L.36.2 BASE-TO-INSTRUCT ORTHOGONAL MAPPING

Within the common PCA subspace, we define

$$Z_B = X_B V_k,$$

and

$$Z_I = X_I V_k.$$

Using only the training samples, we compute

$$M = Z_B^\top Z_I$$

and obtain the orthogonal transformation Q via SVD.

For an evaluation Base activation h_B, we first center it as

$$x = h_B - \mu_B.$$

We decompose the centered activation into its component within the common PCA subspace,

$$x_{\text{sub}} = (xV_k)V_k^\top,$$

and its orthogonal residual component,

$$x_{\text{res}} = x - x_{\text{sub}}.$$

Only the PCA-subspace component is transformed by the Procrustes rotation:

$$z_{\text{aligned}} = (xV_k)Q.$$

The aligned Base-to-Instruct representation is then reconstructed as

$$h_{B\to I}^{\text{aligned}} = z_{\text{aligned}} V_k^\top + x_{\text{res}} + \mu_I.$$

Thus, the Base representation is rotated toward the Instruct coordinate system within the shared low-rank subspace, while the residual component outside the PCA subspace is preserved. We patch this aligned representation into the Instruct Native-Chat computation. This condition is not treated as Primary; rather, it serves as a mechanistic control for whether coordinate remapping can account for failures of direct cross-model activation transfer.

L.37 RECOVERY RATIO

For stimulus i and layer l, we define the distance between the clean Instruct distribution and the Base target distribution as

$$D_{\text{clean},i} = EMD_{VA}\left(P_{I,i}, P_{B,i}\right).$$

Let

$$P_{\text{patch},i,l}$$

denote the output distribution after patching the Base activation at layer l. We define its distance from the Base target distribution as

$$D_{\text{patch},i,l} = EMD_{VA}\left(P_{\text{patch},i,l}, P_{B,i}\right).$$

The Recovery ratio is defined as

$$Recovery_{i,l} = \frac{D_{\text{clean},i} - D_{\text{patch},i,l}}{D_{\text{clean},i} + 10^{-12}}.$$

Under this definition,

$$Recovery > 0$$

indicates that patching moves the Instruct output distribution closer to the Base target distribution. A value of

$$Recovery = 0$$

indicates no improvement in distributional distance, whereas

$$Recovery < 0$$

indicates that patching moves the output distribution farther from the Base target. If the patched distribution exactly matches the Base target distribution, then

$$Recovery = 1.$$

L.38 Layer-Wise Recovery Profile

For each layer, we average Recovery across the evaluation stimuli:

$$\overline{Recovery}(l) = \frac{1}{N_{\text{eval}}} \sum_i Recovery_{i,l}.$$

For each layer, we compute a 95% confidence interval using a stimulus-level bootstrap with

$$B = 1000$$

resamples.

L.39 Primary Depth-Integrated Recovery

Using only the best-performing layer as the Primary outcome would make the analysis sensitive to chance layer-specific fluctuations and post hoc peak selection. We therefore define the Primary Matched-Plain metric in RQ4 as the Area Under the Curve (AUC) of the recovery profile over relative depth:

$$AUC^{\text{MP}}_{\text{Recovery}} = \int_0^1 \overline{Recovery}_{\text{MP}}(d)\, dd.$$

In practice, the integral is approximated by trapezoidal integration over the model's relative-depth grid. We additionally compute a sample-level AUC for each evaluation stimulus from its individual layer-wise recovery profile. Accordingly, Matched-Plain AUC Recovery serves as the Primary distributional recovery quantity in V2 RQ4.

L.40 Absolute EMD Improvement

In addition to normalized Recovery, we record the absolute improvement in distributional distance as

$$\Delta EMD^{\text{MP}}_{i,l} = D^{\text{MP}}_{\text{clean},i} - D^{\text{MP}}_{\text{patch},i,l}.$$

We retain both its layer-wise mean and its AUC over relative depth,

$$AUC^{\text{MP}}_{\Delta EMD}.$$

Whereas the Recovery ratio represents relative improvement normalized by the initial Base–Instruct distance, ΔEMD quantifies the absolute distributional improvement.

L.41 Peak Recovery as a Secondary Localization Measure

For each condition, we define the best recovery layer as

$$l^*_{\text{rec}} = \arg\max_l \overline{Recovery}(l),$$

with the corresponding relative depth

$$d^*_{\text{rec}} = \frac{l^*_{\text{rec}}}{L-1}.$$

We also record

$$\max_l \overline{Recovery}(l)$$

as the maximum recovery ratio.

These quantities are treated as Secondary localization measures. Primary inference for RQ4 is based on the depth-integrated AUC rather than the post hoc best-performing layer.

L.42 Reader–Self Recovery Comparison

RQ4 applies the same recovery protocol to both Reader and Self. For the Primary Matched-Plain condition, we compare

$$AUC^{\mathrm{MP}}_{\mathrm{Self}}$$

and

$$AUC^{\mathrm{MP}}_{\mathrm{Reader}},$$

and define

$$\Delta AUC_{\mathrm{Recovery}} = AUC^{\mathrm{MP}}_{\mathrm{Self}} - AUC^{\mathrm{MP}}_{\mathrm{Reader}}$$

as the Primary cross-family summary of task asymmetry.

The Self–Reader difference in maximum recovery ratio is treated as Secondary. The Native-Chat AUC difference is also Secondary, while the Procrustes-aligned AUC difference is retained as a mechanistic control.

L.43 RQ4 Sample-Level Inference

We additionally perform a mixed-effects analysis of sample-level Matched-Plain recovery, including model family as a fixed effect. The Primary AUC analysis uses

$$AUC^{\mathrm{MP}}_{\mathrm{Recovery}} \sim C(\mathrm{family}) + C(\mathrm{task}) + (1|\mathrm{stimulus}).$$

As a Secondary analysis, we evaluate the Matched-Plain recovery ratio at the global best layer using

$$Recovery^{\mathrm{MP}}_{\mathrm{best}} \sim C(\mathrm{family}) + C(\mathrm{task}) + (1|\mathrm{stimulus}).$$

L.44 Interpretive Boundaries of RQ4

High Recovery indicates that a Base-derived hidden representation can causally move the Instruct output distribution toward the corresponding Base distribution. This does not imply that the model possesses an internal emotion or a human-like subjective affective state.

Conversely, low recovery under Direct patching does not imply that affect-relevant information is absent from the Base representation. If the representational coordinate system differs between Base and Instruct, direct transplantation itself may place the Base activation off the Instruct representational manifold. We therefore include the Procrustes-aligned condition as a mechanistic control for coordinate remapping.

L.45 Four-Family Confirmatory Integration

After completing RQ1–RQ4 separately for each model family, we conduct an integrated confirmatory analysis across the four families in the Primary cohort:

$$\{\mathrm{Qwen}, \mathrm{Llama}, \mathrm{Gemma}, \mathrm{OLMo}\}.$$

L.46 H1a: Geometric Reorganization

For the Base–Instruct Procrustes disparity and RSA under the Matched-Plain condition, we first compute layer-averaged quantities within each model family. For Reader and Self separately, this yields

$$\overline{\mathcal{D}}_{\mathrm{Proc}}, \qquad \overline{\rho}_{\mathrm{RSA}}.$$

Across the four model families, we compute bootstrap confidence intervals using

$$B = 1000$$

resamples. These quantities are treated as cross-family descriptive summaries within the Primary cohort.

L.47 H1B: DECODABILITY PEAK REORGANIZATION

Under the Matched-Plain condition, we compute the family-specific difference in relative depth between the Base and Instruct decodability peaks:

$$\Delta d^*_{D,T} = d^*_{D,I,T} - d^*_{D,B,T}.$$

For Reader and Self, and separately for Valence and Arousal, we report the mean shift across the four model families together with a bootstrap 95% confidence interval. The corresponding peak shift under the Native-Chat condition is treated as Secondary.

L.48 H2: READER–SELF SHARING REORGANIZATION

For each model family, we average the layer-wise Matched-Plain sharing difference,

$$\Delta S_{\mathrm{MP}}(l),$$

across layers to obtain the family-level quantity

$$\overline{\Delta S}_{\mathrm{MP}}.$$

We then report the mean across the four model families together with a bootstrap 95% confidence interval. The corresponding quantity based on Base Plain versus Instruct Native-Chat,

$$\overline{\Delta S}_{\mathrm{Native}},$$

is treated as Secondary.

L.49 H3: CAUSAL-PROFILE REORGANIZATION

Primary inference for H3 is based on the sample-level LMM described above:

$$C_{\mathrm{net,rand}} \sim C(\mathrm{family}) + C(\mathrm{alignment}) \times C(\mathrm{task}) \times d + (1|\mathrm{stimulus}).$$

The Primary condition is Matched-Plain. Benjamini–Hochberg FDR correction is applied to the prespecified interaction terms across Valence and Arousal.

L.50 H4: DISTRIBUTION-RECOVERY ASYMMETRY

The Primary family-level quantity for H4 is

$$AUC^{\mathrm{MP}}_{\mathrm{Self}} - AUC^{\mathrm{MP}}_{\mathrm{Reader}}.$$

Across the four model families, we report the mean difference together with a bootstrap 95% confidence interval. The Native-Chat AUC difference is treated as Secondary, whereas the Procrustes-aligned AUC difference is retained as a mechanistic control. The difference in maximum recovery is also retained as a Secondary localization summary. For sample-level AUC, we use the mixed-effects analysis described above.

L.51 INTERPRETIVE SCOPE OF THE FAMILY-LEVEL BOOTSTRAP

Because the Primary cohort consists of four model families, the number of independent families available for the family-level bootstrap is

$$N_{\mathrm{family}} = 4.$$

Accordingly, the family-level bootstrap intervals are not intended to support strong generalization to an unrestricted population of model families. Rather, they provide a descriptive summary of cross-family consistency among the prespecified Primary families. This level of inference is distinguished from the RQ3 and RQ4 mixed-effects analyses, which leverage sample-level variation.

L.52 H3: Causal-Profile Reorganization

Primary inference for H3 is based on the sample-level LMM described above:

$$C_{\text{net,rand}} \sim C(\text{family}) + C(\text{alignment}) \times C(\text{task}) \times d + (1|\text{stimulus}).$$

The Primary condition is Matched-Plain. Benjamini–Hochberg FDR correction is applied to the prespecified interaction terms across Valence and Arousal.

L.53 H4: Distribution-Recovery Asymmetry

The Primary family-level quantity for H4 is

$$AUC^{\text{MP}}_{\text{Self}} - AUC^{\text{MP}}_{\text{Reader}}.$$

Across the four model families, we report the mean difference together with a bootstrap 95% confidence interval. The Native-Chat AUC difference is treated as Secondary, whereas the Procrustes-aligned AUC difference is retained as a mechanistic control. The difference in maximum recovery is also retained as a Secondary localization summary. For sample-level AUC, we use the mixed-effects analysis described above.

L.54 Interpretive Scope of the Family-Level Bootstrap

Because the Primary cohort consists of four model families, the number of independent families available for the family-level bootstrap is

$$N_{\text{family}} = 4.$$

Accordingly, the family-level bootstrap intervals are not intended to support strong generalization to an unrestricted population of model families. Rather, they provide a descriptive summary of cross-family consistency among the prespecified Primary families. This level of inference is distinguished from the RQ3 and RQ4 mixed-effects analyses, which leverage sample-level variation.

L.55 Organization of Primary and Secondary Evidence in V2

The Primary and major Secondary analyses in V2 are summarized in Table 25.

Table 25: Primary evidence and major control / Secondary analyses in V2.

Analysis	**Primary**	**Secondary / Control**
RQ1 Geometry	Base Plain vs. Instruct Matched-Plain: decodability, cosine RSA, held-out Procrustes disparity	Base vs. Instruct Native-Chat; Instruct Native-Chat vs. Matched-Plain prompt-format effect
RQ2 Sharing	Matched-Plain $\Delta S(l)$ and layer-mean $\overline{\Delta S}$	Native-Chat sharing difference; directional cross-decoding asymmetry
RQ3 Causal Map	Matched-Plain $C_{\text{net,rand}}$; positive-guarded peak / center of mass; sample-level LMM	$C_{\text{net},\perp}$; raw causal displacement; zero ablation; Native-Chat condition
RQ4 Recovery	Matched-Plain depth-integrated AUC Recovery	Matched-Plain $AUC_{\Delta EMD}$; peak recovery; Native-Chat recovery; Procrustes-aligned recovery

M V2 Results

N V2 Discussion

V2 examined whether the affect-relevant representations identified in V1 are simply lost in the transition from Base to Instruct models, or whether the relationship among representational geometry, Reader–Self sharing, causal leverage, and readout into the output distribution is reorganized.

Table 26: V2 H1–H2 (Geometric Reorganization and Changes in Representational Sharing): Post-training-associated representational geometric distortion (Procrustes Distortion), shifts in peak decoding depth Δd^*, and changes in cross-task representational sharing ΔSharing. We report 95% bootstrap confidence intervals aggregated across the four model families.

Hypothesis	Metric	Estimate	95% CI	Condition
H1a: Geometric Distortion	Reader Procrustes Distortion	1.455	[0.639, 2.813]	Matched-Plain
	Self Procrustes Distortion	2.872	[0.660, 6.995]	Matched-Plain
H1b: Decodability Peak Shift	Valence Reader Δd^*	-0.119	$[-0.219, -0.033]$	Matched-Plain
	Valence Self Δd^*	0.007	$[-0.053, 0.100]$	Matched-Plain
	Arousal Reader Δd^*	0.200	$[-0.027, 0.426]$	Matched-Plain
	Arousal Self Δd^*	0.017	$[-0.200, 0.250]$	Matched-Plain
H2: Sharing Reorganization	ΔSharing (Valence)	-0.553	$[-1.324, 0.076]$	Matched-Plain
	ΔSharing (Arousal)	-0.278	$[-0.615, 0.049]$	Matched-Plain

Note: $\Delta d^* = d^*_{\text{instruct}} - d^*_{\text{base}} > 0$ indicates a post-training-associated shift of the representational decodability peak toward later (deeper) layers. Likewise, ΔSharing < 0 indicates a post-training-associated decrease in Reader–Self representational sharing, consistent with reorganization toward greater task separation. These Base–Instruct differences are observational and should not be interpreted as causal effects of post-training itself.

Table 27: V2 H3a (Relocation of Causal Leverage): Post-training-associated shifts in the relative depth of peak causal leverage d^*_C and the causal center-of-mass depth d_{center}, quantified as Δd^*_C and Δd_{center}.

Family	Task	Axis	Base Peak (d^*)	Instruct Peak (d^*)	Δd^*_C	Δd_{center}
GEMMA	Reader	Arousal	0.760	0.560	-0.200	-0.106
GEMMA	Reader	Valence	0.520	0.080	-0.440	0.098
GEMMA	Self	Arousal	0.560	0.240	-0.320	0.010
GEMMA	Self	Valence	0.000	0.520	0.520	0.044
LLAMA	Reader	Arousal	0.467	0.467	0.000	0.013
LLAMA	Reader	Valence	0.533	0.000	-0.533	-0.089
LLAMA	Self	Arousal	0.067	0.800	0.733	0.254
LLAMA	Self	Valence	0.467	0.933	0.467	0.341
OLMO	Reader	Arousal	0.200	0.600	0.400	0.043
OLMO	Reader	Valence	0.600	0.267	-0.333	-0.170
OLMO	Self	Arousal	0.067	0.333	0.267	-0.087
OLMO	Self	Valence	0.600	0.133	-0.467	-0.296
QWEN	Reader	Arousal	0.037	0.000	-0.037	-0.144
QWEN	Reader	Valence	0.778	0.111	-0.667	-0.042
QWEN	Self	Arousal	0.296	0.852	0.556	0.027
QWEN	Self	Valence	0.593	0.259	-0.333	-0.178

Note: The shifts in causal peak and center-of-mass depth, $\Delta d^*_C = d^*_{C,\text{Instruct}} - d^*_{C,\text{Base}}$ and Δd_{center}, quantify post-training-associated changes in the localization of causal leverage toward later layers. This aggregate table is currently unevaluated (—) and is therefore not used as evidence for causal relocation or related conclusions. Statistical inference for H3 instead relies on the sample-level LMM reported in Table 29. Because the Base–Instruct comparison is not a randomized intervention on post-training, these differences are interpreted as post-training-associated reorganization rather than as causal effects of post-training itself.

Table 28: V2 H3b (Causal Specificity Controls): Raw causal effect C_{raw}, random-direction control C_{rand}, orthogonal-direction control C_{perp}, net causal effect $C_{\text{net,rand}} = C_{\text{raw}} - C_{\text{rand}}$, and zero-ablation effect.

Family	Task	Condition	Axis	C_{raw}	C_{rand}	C_{perp}	$C_{\text{net,rand}}$ (**Primary**)	Zero-Ablation
OLMO	Reader	Base-Plain	Valence	0.002	0.002	0.002	−0.000	0.016
OLMO	Reader	Base-Plain	Arousal	0.001	0.001	0.001	0.000	0.019
OLMO	Self	Base-Plain	Valence	0.002	0.002	0.002	−0.000	0.016
OLMO	Self	Base-Plain	Arousal	0.001	0.001	0.001	0.000	0.020
OLMO	Reader	Matched-Plain	Valence	0.004	0.004	0.004	−0.000	0.053
OLMO	Reader	Matched-Plain	Arousal	0.001	0.001	0.001	−0.000	0.016
OLMO	Self	Matched-Plain	Valence	0.004	0.004	0.004	−0.000	0.075
OLMO	Self	Matched-Plain	Arousal	0.001	0.001	0.001	0.000	0.037
LLAMA	Reader	Base-Plain	Valence	0.003	0.003	0.003	0.000	0.033
LLAMA	Reader	Base-Plain	Arousal	0.001	0.001	0.001	−0.000	0.015
LLAMA	Self	Base-Plain	Valence	0.003	0.003	0.003	0.000	0.032
LLAMA	Self	Base-Plain	Arousal	0.001	0.001	0.001	−0.000	0.014
LLAMA	Reader	Matched-Plain	Valence	0.003	0.003	0.003	−0.000	0.020
LLAMA	Reader	Matched-Plain	Arousal	0.001	0.001	0.001	−0.000	0.011
LLAMA	Self	Matched-Plain	Valence	0.003	0.003	0.003	−0.000	0.020
LLAMA	Self	Matched-Plain	Arousal	0.001	0.001	0.001	−0.000	0.012
GEMMA	Reader	Base-Plain	Valence	0.005	0.005	0.005	0.000	0.031
GEMMA	Reader	Base-Plain	Arousal	0.001	0.001	0.001	0.000	0.051
GEMMA	Self	Base-Plain	Valence	0.005	0.005	0.005	−0.000	0.036
GEMMA	Self	Base-Plain	Arousal	0.001	0.001	0.001	−0.000	0.054
GEMMA	Reader	Matched-Plain	Valence	0.006	0.006	0.006	−0.000	0.054
GEMMA	Reader	Matched-Plain	Arousal	0.001	0.001	0.001	0.000	0.036
GEMMA	Self	Matched-Plain	Valence	0.006	0.006	0.006	0.000	0.059
GEMMA	Self	Matched-Plain	Arousal	0.001	0.001	0.001	0.000	0.030
QWEN	Reader	Base-Plain	Valence	0.004	0.004	0.004	0.000	0.029
QWEN	Reader	Base-Plain	Arousal	0.001	0.001	0.001	0.000	0.008
QWEN	Self	Base-Plain	Valence	0.004	0.004	0.004	−0.000	0.022
QWEN	Self	Base-Plain	Arousal	0.001	0.001	0.001	−0.000	0.007
QWEN	Reader	Matched-Plain	Valence	0.007	0.007	0.007	−0.000	0.122
QWEN	Reader	Matched-Plain	Arousal	0.001	0.001	0.001	0.000	0.028
QWEN	Self	Matched-Plain	Valence	0.007	0.007	0.007	0.000	0.121
QWEN	Self	Matched-Plain	Arousal	0.001	0.001	0.001	−0.000	0.027

Note: $C_{\text{net,rand}} > 0$ indicates that the mean intervention effect along the affect-related direction exceeds that of the random-direction control. This aggregate table is currently unevaluated (—) and is therefore not used as the basis for statistical conclusions regarding H3; inference instead relies on the sample-level LMM reported in Table 29. Because the Base–Instruct comparison is not a randomized intervention on post-training, these differences are interpreted as post-training-associated reorganization rather than as causal effects of post-training itself.

Table 29: V2 H3c (Linear Mixed-Effects Model of Causal Reorganization): Sample-level regression / mixed-effects analysis of control-adjusted causal leverage $C_{\text{net,rand}}$ as a function of family, alignment, task, relative depth, and their interactions. We report the prespecified Primary interaction terms and the Secondary post-training main effect for both Valence and Arousal.

Axis	Predictor / Parameter	Estimate (β)	95% CI	p-value	FDR q
Valence	Post-training × Depth (**Primary**)	0.000090	[−0.000054, 0.000234]	0.223	0.304
	Post-training × Task (**Primary**)	0.000163	[0.000044, 0.000282]	0.007	0.044
	Post-training × Task × Depth (**Primary**)	−0.000192	[−0.000396, 0.000012]	0.065	0.196
	Post-training (Instruct = 1) (**Secondary**)	−0.000095	[−0.000179, −0.000011]	0.027	—
Arousal	Post-training × Depth (**Primary**)	−0.000019	[−0.000055, 0.000017]	0.304	0.304
	Post-training × Task (**Primary**)	−0.000017	[−0.000047, 0.000013]	0.277	0.304
	Post-training × Task × Depth (**Primary**)	0.000034	[−0.000017, 0.000085]	0.196	0.304
	Post-training (Instruct = 1) (**Secondary**)	0.000006	[−0.000015, 0.000028]	0.549	—

Note: Primary confirmatory inference is based on the prespecified Alignment × Depth, Alignment × Task, and Alignment × Task × Depth interaction terms. For Valence, the Post-training × Task interaction remained significant after FDR correction ($q = 0.044$), providing partial support for task-dependent post-training-associated reorganization of causal leverage. In contrast, interactions involving depth, including Post-training × Depth, did not meet the FDR-corrected significance criterion, nor did any of the Primary interactions for Arousal. Because the Base–Instruct comparison is not a randomized intervention on post-training, these differences are interpreted as post-training-associated reorganization rather than as causal effects of post-training itself.

Table 30: V2 H4 (Output-Distribution Recovery via Intervention): Recovery of the Instruct model's output distribution through interventions using representations from the corresponding Base model. We report Matched-Plain AUC, recovery in distributional distance (ΔEMD AUC), maximum recovery (Max Recovery), optimal recovery depth (Best Depth), and recovery under the Native and Aligned conditions.

Family	Task	Matched AUC	ΔEMD AUC	Max Recovery	Best Depth (d^*)	Native AUC	Aligned AUC
QWEN	Reader	0.189	0.056	0.439	0.370	0.007	−0.001
QWEN	Self	0.181	0.062	0.417	0.407	0.052	−0.000
LLAMA	Reader	0.005	0.001	0.113	0.400	0.158	0.001
LLAMA	Self	−0.006	−0.000	0.078	0.400	0.125	−0.000
GEMMA	Reader	−0.052	−0.012	0.142	0.040	−0.072	−0.011
GEMMA	Self	0.001	0.005	0.326	0.000	−0.025	−0.009
OLMO	Reader	−0.170	−0.037	0.006	0.800	0.044	0.001
OLMO	Self	−0.160	−0.045	0.000	1.000	0.070	−0.001

Note: We evaluate the extent to which replacing or aligning an Instruct model's internal representations with representations derived from the corresponding Base model moves the Instruct VA output distribution toward the Base distribution, as quantified by Matched AUC, ΔEMD AUC, and Max Recovery. Under the prespecified four-family design (Qwen, Llama, Gemma, and OLMo), Matched-Plain recovery was empirically evaluated for all models. Recovery was strongly family-dependent, with positive recovery for Qwen, near-zero recovery for Llama, and little or negative recovery for Gemma and OLMo in several conditions. Because the Base–Instruct comparison is not a randomized intervention on post-training, these differences are interpreted as post-training-associated reorganization rather than as causal effects of post-training itself.

Table 31: V2 Cross-Family Hypothesis Summary (H1–H4): Summary of results for the prespecified V2 hypotheses, including estimates for the key analysis metrics, cross-family bootstrap 95% confidence intervals, and the corresponding criterion / interpretation.

Hypothesis	Key Analysis Metric	Estimate	95% CI	p / FDR q	Criterion / Interpretation
H1a: Geometry Reorganization	Reader Procrustes Distortion	1.455	[0.639, 2.813]	—	Descriptive
H1a: Geometry Reorganization	Self Procrustes Distortion	2.872	[0.660, 6.995]	—	Descriptive
H1b: Decodability Peak Shift	Valence Reader Peak Shift Delta d*	−0.119	[−0.219, −0.033]	—	Not Supported
H1b: Decodability Peak Shift	Valence Self Peak Shift Delta d*	0.007	[−0.053, 0.100]	—	Not Supported
H1b: Decodability Peak Shift	Arousal Reader Peak Shift Delta d*	0.200	[−0.027, 0.426]	—	Not Supported
H1b: Decodability Peak Shift	Arousal Self Peak Shift Delta d*	0.017	[−0.200, 0.250]	—	Not Supported
H2: Sharing Reorganization	Valence Delta Sharing	−0.553	[−1.324, 0.076]	—	Not Supported
H2: Sharing Reorganization	Arousal Delta Sharing	−0.278	[−0.615, 0.049]	—	Not Supported
H3: Causal Reorganization	Valence Alignment x Depth (Primary)	0.000	[−0.000, 0.000]	0.304	Not Supported
H3: Causal Reorganization	Valence Alignment x Task (Primary)	0.000	[0.000, 0.000]	0.044	✓Supported
H3: Causal Reorganization	Valence Alignment x Task x Depth (Primary)	−0.000	[−0.000, 0.000]	0.196	Not Supported
H3: Alignment Main Effect	Valence Alignment main effect (Secondary)	−0.000	[−0.000, −0.000]	0.027	✓Supported
H3: Causal Reorganization	Arousal Alignment x Depth (Primary)	−0.000	[−0.000, 0.000]	0.304	Not Supported
H3: Causal Reorganization	Arousal Alignment x Task (Primary)	−0.000	[−0.000, 0.000]	0.304	Not Supported
H3: Causal Reorganization	Arousal Alignment x Task x Depth (Primary)	0.000	[−0.000, 0.000]	0.304	Not Supported
H3: Alignment Main Effect	Arousal Alignment main effect (Secondary)	0.000	[−0.000, 0.000]	0.549	Not Supported
H4: Recovery Asymmetry	Self–Reader AUC diff (Primary)	0.011	[−0.010, 0.038]	—	Not Supported
H4: Recovery Asymmetry	Matched-Plain AUC (Descriptive)	−0.001	—	—	—

Note: H1a is treated as a cross-family descriptive summary. Representation-geometric disparity was observed between Base and Instruct models for both Reader and Self, but exclusion of zero by the Procrustes Distortion confidence intervals is not interpreted as a null-hypothesis significance test. In contrast, the prespecified positive peak-shift criterion for H1b and the criterion for Reader–Self sharing reorganization in H2 were not satisfied based on the four-family bootstrap confidence intervals. H3 (Causal Reorganization) was not uniformly unsupported: the sample-level LMM provided partial support for a task-dependent Valence difference (Alignment × Task, FDR $q = 0.044$), whereas depth relocation, including the Alignment × Depth interaction, was not supported. Because the Base–Instruct comparison is not a randomized intervention on post-training, these differences are interpreted as post-training-associated reorganization rather than as causal effects of post-training itself. H4 Recovery Asymmetry, quantified by the Self–Reader AUC difference, was also not supported because its 95% confidence interval included zero.

Overall, the V2 results are more consistent with a change in the internal organization of affect-relevant information while that information remains decodable than with simple representational erasure. At the same time, we found no evidence that this change can be reduced to a uniform shift toward later layers, systematic Reader–Self separation, or a single causal bottleneck.

N.1 Base–Instruct Differences Were Characterized by Representational Geometry Rather Than Information Loss

For H1a, the Base–Instruct Procrustes Distortion in the Matched-Plain condition was

$$1.455, \qquad 95\%\,\mathrm{CI} = [0.639, 2.813]$$

for Reader, and

$$2.872, \qquad 95\%\,\mathrm{CI} = [0.660, 6.995]$$

for Self. Thus, under Matched-Plain prompting, consistent Base–Instruct representational-geometric disparities were observed for both Reader and Self. We treat these estimates as descriptive cross-family summaries within the Primary cohort.

Importantly, V1 showed that information related to human Reader V/A remained decodable on held-out data from both Base and Instruct representations. The Base–Instruct difference is therefore not well characterized by a simple erasure account of the form

$$\text{Base: Information Exists} \quad \rightarrow \quad \text{Instruct: Information Disappears.}$$

Instead, the combined results are more consistent with

$$\text{Affect-Relevant Information Remains Decodable} \quad + \quad \text{Its Representational Geometry Differs.}$$

This distinction suggests that understanding post-training-associated behavioral change requires separating the presence of information from its coordinate organization within hidden-state space. The same externally defined affective variable may remain recoverable while its geometric organization differs between Base and Instruct models.

N.2 Geometric Reorganization Did Not Imply a Uniform Shift Toward Later Layers

The simple relocation hypothesis tested in H1b—that decodability peaks would consistently shift toward deeper layers in Instruct models—was not supported. For Valence Reader,

$$\Delta d_D^* = -0.119, \qquad 95\%\,\mathrm{CI} = [-0.219, -0.033],$$

indicating a shift in the direction opposite to the prespecified deeper-layer hypothesis, with the peak occurring at an earlier relative depth in Instruct models.

For Valence Self,

$$\Delta d_D^* = 0.007, \qquad 95\%\,\mathrm{CI} = [-0.053, 0.100],$$

and for Arousal Self,

$$\Delta d_D^* = 0.017, \qquad 95\%\,\mathrm{CI} = [-0.200, 0.250],$$

both estimates were close to zero. For Arousal Reader,

$$\Delta d_D^* = 0.200, \qquad 95\%\,\mathrm{CI} = [-0.027, 0.426],$$

the confidence interval also included zero. Thus,

$$\text{Geometric Reorganization} \quad \neq \quad \text{Uniform Deeper Relocation.}$$

Although Base and Instruct models differ in their representational geometry, these differences do not manifest as a uniform shift of affect-relevant decodability toward later layers across tasks and affective axes.

N.3 A Consistent Reduction in Reader–Self Sharing Was Not Supported

For H2, the change in Reader–Self representational sharing under the Matched-Plain condition was

$$\Delta\text{Sharing} = -0.553, \qquad 95\%\,\text{CI} = [-1.324, 0.076]$$

for Valence, and

$$\Delta\text{Sharing} = -0.278, \qquad 95\%\,\text{CI} = [-0.615, 0.049]$$

for Arousal. Although both point estimates were in the direction of greater task separation, both 95% confidence intervals included zero.

H2 therefore did not support a consistent cross-family separation of Reader and Self affect-relevant representations associated with the Base–Instruct difference. Moreover, in the Secondary Native Chat condition, differences in prompt format substantially amplified the apparent reduction in sharing. This result underscores the importance of the Matched-Plain control for separating Base–Instruct representational differences from prompt-format effects. In other words,

$$\text{Geometric Reorganization} \quad \not\Rightarrow \quad \text{Reader–Self Sharing Uniformly Decreases.}$$

N.4 Evidence for Causal Reorganization Was Limited to a Task-Dependent Valence Component

In RQ3, we evaluated differences in causal profiles between Base and Instruct models using a sample-level LMM on the control-adjusted causal leverage $c_{\text{net,rand}}$. For Valence, the Post-training $\times$ Task interaction was

$$\beta = 1.63 \times 10^{-4}, \qquad \text{FDR } q = 0.044,$$

making it the only Primary interaction that remained significant after FDR correction. The Secondary Post-training main effect was also significant ($q = 0.027$).

By contrast, neither the Post-training $\times$ Depth interaction (FDR $q = 0.304$) nor the Post-training $\times$ Task $\times$ Depth interaction (FDR $q = 0.196$) was supported. None of the corresponding Arousal interaction terms survived FDR correction (FDR $q = 0.304$).

The H3 results therefore support neither complete invariance of causal organization nor a uniform relocation toward deeper layers. More narrowly, Valence showed evidence of a task-dependent change in the causal profile associated with Base–Instruct alignment, whereas a consistent cross-family relocation of causal leverage across depth was not statistically supported.

N.5 Recovery Through Base-Representation Patching Was Strongly Family-Dependent

RQ4 tested whether replacing or aligning an Instruct model's internal representation with a representation derived from the corresponding Base model moved the Instruct VA output distribution toward the Base distribution. The results differed substantially across model families.

For Qwen, the Matched-Plain Recovery AUC reached 0.189 for Reader and 0.181 for Self, with maximum Recovery of approximately 0.42–0.44. In contrast, the Matched-Plain AUC was close to zero for Llama (Reader 0.005, Self -0.006) and was negative or approximately zero for Gemma (Reader -0.052, Self 0.001) and OLMo (Reader -0.170, Self -0.160). Thus,

$$\text{Base Representation Patching} \quad \not\Rightarrow \quad \text{Universal Base-Like Recovery.}$$

Furthermore, the Primary H4 Self–Reader recovery asymmetry was

$$\Delta\text{AUC}_{\text{Recovery}} = 0.011, \qquad 95\%\,\text{CI} = [-0.010, 0.038],$$

with the confidence interval including zero. The hypothesis that Self representations are consistently more recoverable toward the Base-like distribution than Reader representations was therefore not supported. Procrustes-aligned AUCs were also close to zero across families, indicating that the Base–Instruct discrepancy cannot be reduced to a simple linear coordinate mismatch.

N.6 Integrated Interpretation of V2

Taken together, the V2 results cannot be reduced to any single account based on Erasure, Uniform Deeper Relocation, Uniform Reader–Self Separation, or a Single Recoverable Bottleneck.

Under Matched-Plain prompting, consistent Base–Instruct representational-geometric disparities were observed for both Reader and Self, while affect-relevant information remained decodable in Instruct models. These differences, however, did not manifest as a uniform shift of decodability peaks toward later layers or as a consistent reduction in Reader–Self sharing.

At the causal level, the Valence Post-training $\times$ Task interaction satisfied the Primary criterion, whereas depth-dependent relocation was not supported. Recovery of the Base-like output distribution through Base-representation patching was also strongly model-family-dependent.

Overall, these findings are consistent with post-training-associated reorganization distributed across representational geometry, task-dependent causal sensitivity, and model-specific recoverability, rather than with a single localized mechanism. Because the Base and Instruct models are separately obtained model artifacts rather than randomized experimental conditions produced by a controlled post-training intervention, however, these differences should be interpreted as post-training-associated reorganization rather than as causal effects of post-training itself.

O From V2 to V3

V2 revealed clear Base–Instruct differences in representational geometry, as measured by Procrustes Distortion. However, these differences could not be explained by a consistent shift of decodability peaks toward later layers, reduced Reader–Self sharing, relocation of causal leverage across depth, or simple linear transplantation of Base representations.

In particular, the Post-training $\times$ Task interaction received partial support for Valence, whereas interactions involving layer depth were not supported. Thus, understanding post-training-associated change requires more than a static layer-wise comparison of where affect-relevant information is represented.

The internal representations and causal profiles examined in V2 were primarily measured at the end of prompt processing, using prompt-end hidden states. Self-Report, however, is not determined instantaneously at the prompt boundary. Rather, it emerges through a dynamic autoregressive computation as the model generates the response token sequence. The next stage therefore extends the analysis from a static layer axis to the two-dimensional spatiotemporal coordinate system

$$\text{Layer} \times \text{Generation Stage}.$$

Moreover, as illustrated by V2 RQ4, the presence of an internal representation is not equivalent to its ability to robustly exert causal control over the output distribution. A further distinction is required between artificial sensitivity—the ability of an external intervention to move model outputs—and endogenous relevance—whether the corresponding representation is naturally utilized during the model's forward computation.

V3 therefore extends the analysis of Self-Report computation in Instruct models from static prompt-end representations to the response-generation trajectory. Specifically, we first apply a prespecified State Induction Gate based on Sufficiency, Specificity, and Endogenous Relevance to test whether affect-related directions exhibit measurable causal leverage. We then analyze causal sensitivity jointly across layer (space) and generation stage (time) along a teacher-forced trajectory defined by a fixed canonical response candidate.

Accordingly, the generation-stage analysis in V3 does not characterize the model's unconstrained, freely generated response trajectory. Instead, it provides a controlled spatiotemporal comparison along a fixed trajectory in which candidate content is held constant.

P V3 Overview

The outline of V3 stage is shown in Fig. 5.

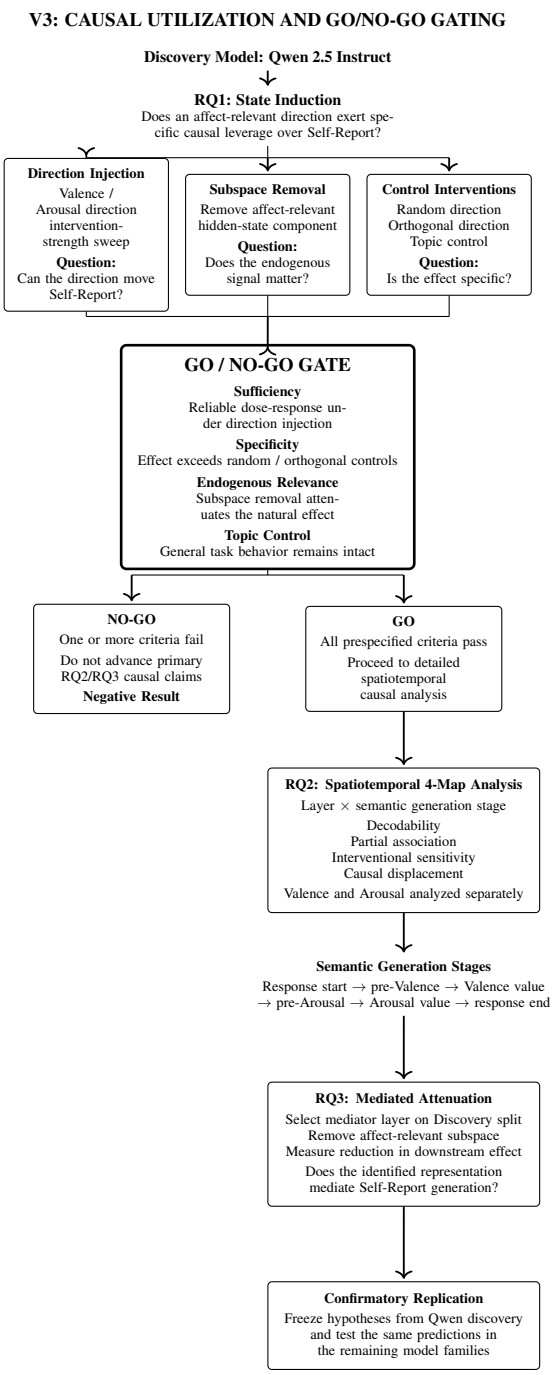


Figure 5: Overview of V3. RQ1 tests whether affect-relevant directions exert specific causal leverage over Self-Report. The Go/No-Go gate evaluates sufficiency, specificity, endogenous relevance, and preservation of general task behavior. If any criterion fails, the pipeline stops and the outcome is retained as a negative result. If all criteria pass, RQ2 maps effects across layers and semantic generation stages, followed by RQ3 mediated-attenuation analysis and confirmatory replication.

P.1 Objective and Positioning

The central Research Question in V3 is: "At which layers and generation stages does affect-relevant information exert measurable causal leverage over Self-Report in Instruct models?"

A critical distinction throughout V3 is

$$\text{Information is decodable} \quad \neq \quad \text{Information has causal leverage.}$$

The ability to decode affective information accurately from an internal representation does not imply that the decoded information causally drives output generation.

V3 first establishes the necessary preconditions through a State Induction Gate, then maps the spatiotemporal causal structure within Instruct models, and finally evaluates reproducibility on an independent confirmation set and across model families:

$$\begin{gathered}\text{State Induction / Sufficiency} \\ \downarrow \\ \text{Specificity and Endogenous Relevance} \\ \downarrow \\ \text{Spatiotemporal Causal Mapping} \\ \downarrow \\ \text{Mediated Attenuation} \\ \downarrow \\ \text{Cross-Family Confirmation.}\end{gathered}$$

P.2 Overall Structure of V3

V3 consists of a Discovery analysis using Qwen 2.5 1.5B-Instruct followed by Confirmatory replication in the remaining model families. The Discovery model is Qwen 2.5 1.5B-Instruct. The Confirmatory families are

$$\{\text{Llama 3.2 Instruct}, \text{Gemma 3 Instruct}, \text{OLMo 2 Instruct}\}.$$

Base models are not included in V3. Table 32 summarizes the analyses comprising V3.

Table 32: Discovery and Confirmatory analyses in V3.

Stage	Analysis	Question
RQ1	State Induction & Go/No-Go Gate	Does intervention along an affect-related direction alter Self-Report in a dose-dependent and specific manner, and does removing the same subspace during natural stimuli attenuate the corresponding Self shift?
RQ2	Spatiotemporal 4-Maps	How are decodability, endogenous relevance, interventional slope, and causal displacement distributed across layer $\times$ generation stage?
RQ3	Mediated Attenuation	To what extent does removing the affect-related subspace selected during Discovery attenuate the natural Self shift on independent Confirmation pairs?
Confirmation	Cross-Family Replication	When the site / stage selected in Qwen is held fixed, is the same mechanistic pattern observed in Llama, Gemma, and OLMo?

P.3 V3 Data and Measurement Targets

V3 uses only the

$$192 \text{ Clinical–Neutral matched pairs}$$

from AIPsy-Affect. Each analysis unit consists of a Clinical stimulus and its matched Neutral stimulus sharing the same `pair_id`:

$$(x_i^C, x_i^N), \qquad i = 1, \ldots, 192.$$

Moderate and Complex Neutral stimuli are not used in the V3 Primary analyses. Every Clinical observation is required to have its corresponding matched Neutral text. The Primary production analyses do not use fallback substitutes such as a generic neutral sentence or a fixed value of 5.0.

P.4 Primary Affective Grounding

Because AIPsy-Affect does not provide continuous human Valence / Arousal annotations analogous to those available in EmoBank, we do not treat nonexistent human V/A labels as ground truth. Instead, for each stimulus, the Primary affective targets are obtained by evaluating the same Instruct model under the Reader task:

$$R_V(x)$$

and

$$R_A(x).$$

Accordingly,

$$Y_V = R_V(x), \qquad Y_A = R_A(x).$$

These targets are not human affective ground truth; they are

model Reader Prediction.

V3 measures the extent to which Reader-grounded affective information is represented in Self hidden states and how manipulating that information alters Self-Report. Thus, the central mechanistic relationship examined in V3 is

Reader-grounded Representation $\rightarrow$ Self-Report Causal Utilization.

Directions derived using Self-Report itself as the target are retained only as a Secondary characterization.

P.5 V3 Prompts

The Reader and Self task instructions and the 81-state VA JSON output schema follow exactly the common prompt specification defined in §C.6.2. Because V3 considers only Instruct models, we consistently use each model's native chat format throughout V3.

P.6 Topic Control Task

In RQ1, we test whether intervention along an affect-related direction alters Self-Report without acting merely as a generic perturbation that substantially disrupts an arbitrary non-affective task distribution. To this end, we use the following topic-classification task on the matched Neutral texts:

```
Classify the primary topic of the following text
into one of: 'medical', 'family', 'work', 'daily_life'.

Respond strictly in JSON format:
{"topic": "<medical|family|work|daily_life>"}
```

The candidate set consists of the four categories

$$\{\texttt{medical}, \texttt{family}, \texttt{work}, \texttt{daily_life}\}.$$

P.7 RQ1: State Induction and Go/No-Go Gate

Before proceeding to the subsequent spatiotemporal causal analyses, RQ1 tests whether intervention along a Reader-grounded affect direction exerts sufficient causal leverage over Self-Report. The criterion is not merely whether injecting the direction changes the output in some manner. Instead, we jointly require the following:

1. **Sufficiency / Dose–Response**: Self-Report changes systematically as a function of intervention dose.
2. **Specificity**: Intervention along the affect-related direction produces a larger effect than interventions along random or orthogonal directions.
3. **Endogenous Relevance**: Removing the affect-related subspace during natural affective stimulus processing attenuates the Clinical–Neutral Self shift.
4. **Topic Selectivity Control**: The same intervention does not substantially disrupt the distribution of a non-affective topic-classification task.

P.8 RQ1 TRAIN–TEST SPLIT

We shuffle the 192 matched pairs at the `pair_id` level using seed 42 and split them into

$$70\% \text{ train} + 30\% \text{ held-out test.}$$

For the current 192-pair dataset, this yields

$$N_{\text{train}} = 134, \qquad N_{\text{test}} = 58.$$

No pair appears in both the training and test subsets. The affect direction, control directions, neutral mean representation, and other quantities required for the interventions are estimated exclusively from the training subset. Intervention outcomes used for the Go/No-Go decision are evaluated exclusively on the held-out test pairs.

P.9 RQ1 TARGET SITE

RQ1 does not select the layer with the largest effect after inspecting the subsequent spatiotemporal causal maps. Instead, the target relative depth is prespecified as

$$d = 0.5.$$

For a model with L Transformer blocks, the target layer is defined as

$$l_{\text{RQ1}} = \text{round}\left[0.5(L-1)\right].$$

The intervention site is the prompt-end post-MLP residual state under the Self prompt.

P.10 READER-GROUNDED CONDITIONAL DIRECTIONS

For the training Clinical stimuli, let

$$H \in \mathbb{R}^{N_{\text{train}} \times D}$$

denote the Self representations at the target layer. For the same stimuli, let the corresponding Reader Predictions be

$$V_R \in \mathbb{R}^{N_{\text{train}}},$$

and

$$A_R \in \mathbb{R}^{N_{\text{train}}}.$$

In contrast to a conventional probe that predicts V/A from hidden representations, RQ1 performs a multi-output regression in the reverse direction, using Valence and Arousal as predictors of the hidden representation:

$$H = \mathbf{1}b^\top + V_R\beta_V^\top + A_R\beta_A^\top + E.$$

The Ridge regularization parameter is fixed to

$$\alpha_{\text{Ridge}} = 1.$$

This joint regression yields a Valence-associated hidden direction,

$$\beta_V,$$

and an Arousal-associated hidden direction,

$$\beta_A,$$

with each coefficient estimated conditional on the other affective axis. We define the corresponding unit directions as

$$d_V = \frac{\beta_V}{\|\beta_V\|_2},$$

and

$$d_A = \frac{\beta_A}{\|\beta_A\|_2}.$$

As a Secondary characterization, we apply the same procedure to the training representations using Self-Report V/A as the regression targets. We then quantify the alignment between the Reader-grounded and Self-derived directions as

$$\cos(d_V^R, d_V^S), \qquad \cos(d_A^R, d_A^S).$$

P.11 Dose Normalization

For a direction d, we project the training representations onto that direction as

$$z_i = h_i^\top d.$$

We define the standard deviation of the directional activation as

$$\sigma_d = \mathrm{SD}(Hd).$$

The intervention is applied as

$$h' = h + \alpha \sigma_d d.$$

Thus, $\alpha = 1$ corresponds to an intervention magnitude of one standard deviation of directional activation in the training representations.

P.12 RQ1 Sufficiency / Dose–Response

For each held-out pair, we inject either the Valence or Arousal direction into the Self representation of the matched Neutral stimulus. The intervention dose grid is

$$\alpha \in \{-1, -0.5, 0, 0.5, 1\}.$$

For the Valence direction, let

$$V_i^N$$

denote the clean Self expectation for the Neutral stimulus, and let

$$V_i^{N,\alpha}$$

denote the corresponding value under intervention dose α. We define the intervention-induced shift as

$$\Delta V_i(\alpha) = V_i^{N,\alpha} - V_i^N.$$

Analogously, for Arousal,

$$\Delta A_i(\alpha) = A_i^{N,\alpha} - A_i^N.$$

For each held-out pair, we estimate the linear slope relating intervention dose to the corresponding report shift:

$$\gamma_i = \frac{\sum_\alpha (\alpha - \bar{\alpha})(\Delta Y_i(\alpha) - \overline{\Delta Y_i})}{\sum_\alpha (\alpha - \bar{\alpha})^2}.$$

For Valence and Arousal separately, we bootstrap the sample-level slopes and report the mean slope together with its 95% confidence interval.

P.13 RQ1 Specificity Controls

To test whether the effect of intervention along an affect-related direction can be explained by a generic consequence of hidden-state perturbation, we generate, for each affective axis,

$$K = 5$$

random control directions and

$$K = 5$$

orthogonal control directions.

Each random direction is normalized to have the same norm as the target direction. Each orthogonal direction is constructed by removing the component parallel to the target direction from a random vector and then renormalizing the resulting vector to the same norm as the target direction. The reference intervention dose is

$$\alpha = 1.$$

For pair i, we define the absolute effect of the affect-direction intervention as

$$E_{i,\text{aff}} = |\Delta Y_{i,\text{aff}}| .$$

Let

$$\overline{E}_{i,\text{rand}}$$

denote the mean absolute effect across the five random directions, and let

$$\overline{E}_{i,\perp}$$

denote the corresponding mean across the five orthogonal directions.

We define the Primary specificity measure as

$$S_i = E_{i,\text{aff}} - \max\left(\overline{E}_{i,\text{rand}}, \overline{E}_{i,\perp}\right).$$

For Valence and Arousal separately, we bootstrap the sample-level specificity scores and compute 95% confidence intervals. We additionally retain

$$E_{\text{aff}} - \overline{E}_{\text{rand}}$$

and

$$E_{\text{aff}} - \overline{E}_{\perp}$$

as Secondary diagnostics relative to the random and orthogonal controls, respectively.

P.14 RQ1 Endogenous Relevance

The ability to alter Self-Report through directional injection establishes sufficiency, but does not establish that the corresponding representation is endogenously utilized during natural Clinical stimulus processing. We therefore construct an affect-related subspace from the Reader-grounded Valence and Arousal directions:

$$B = [d_V \;\; d_A].$$

We use SVD to remove rank-deficient directions and obtain an orthonormal basis

$$Q \in \mathbb{R}^{D\times r}, \qquad r \le 2,$$

satisfying

$$Q^\top Q = I_r.$$

Let

$$\mu_{\text{neu}}$$

denote the mean Self representation of the matched Neutral stimuli in the training pairs. For a held-out Clinical representation h_i^C, we define the centered representation as

$$h_{i,\text{centered}} = h_i^C - \mu_{\text{neu}}.$$

We then remove the affect-related subspace component

$$QQ^\top\left(h_i^C - \mu_{\text{neu}}\right)$$

from the Clinical representation:

$$h_i' = h_i^C - QQ^\top\left(h_i^C - \mu_{\text{neu}}\right).$$

P.15 Natural and Residual Self Shifts

For each held-out pair i, we define the difference between the clean Clinical Self output and the matched Neutral Self output as

$$N_{i,V} = \left|V_i^C - V_i^N\right|,$$

and

$$N_{i,A} = \left|A_i^C - A_i^N\right|.$$

We refer to these quantities as the natural affective shifts.

Let $V_i^{C,\mathrm{abl}}$ and $A_i^{C,\mathrm{abl}}$ denote the Clinical outputs after affect-related subspace removal. We define the corresponding residual shifts as

$$R_{i,V} = \left| V_i^{C,\mathrm{abl}} - V_i^N \right|,$$

and

$$R_{i,A} = \left| A_i^{C,\mathrm{abl}} - A_i^N \right|.$$

The absolute attenuation is defined as

$$N_i - R_i.$$

For the RQ1 Gate, we additionally use a scale-normalized attenuation ratio. To avoid instability when the natural shift is near zero, this ratio is evaluated only for samples satisfying

$$N_i > 0.05.$$

For these samples, we define

$$AR_i = \frac{N_i - R_i}{N_i}.$$

For Valence and Arousal separately, we bootstrap the distribution of sample-level attenuation ratios.

P.16 RQ1 Topic Control

We inject the same Reader-grounded directions into the topic-classification computation for the matched Neutral texts. Let

$$P$$

denote the clean topic distribution and

$$P'$$

the corresponding distribution after intervention. We quantify the distributional change using total variation distance:

$$TVD(P, P') = \frac{1}{2} \sum_k \left| P'_k - P_k \right|.$$

The Valence and Arousal directions are evaluated separately. Importantly, Topic TVD and the Self VA shift quantify different constructs. We therefore do not define a quantity such as

$$\text{Self Effect} - \text{Topic TVD}$$

as a Primary causal effect. Instead, Topic TVD is treated as an independent control for assessing whether the intervention induces substantial non-specific disruption of an unrelated task distribution.

P.17 RQ1 Bootstrap and Gate Criteria

Primary confidence intervals in RQ1 are estimated using

$$B = 1000$$

bootstrap resamples. For each of the Valence and Arousal axes, the following four criteria are evaluated independently.

Criterion 1: Sufficiency. We require the lower bound of the 95% confidence interval for the interventional slope to satisfy

$$CI_{\mathrm{low}}(\gamma) > 0.1.$$

Criterion 2: Specificity. We require the lower bound of the 95% confidence interval for the specificity score to satisfy

$$CI_{\mathrm{low}}(S) > 0.05.$$

Criterion 3: Endogenous Relevance. We require the lower bound of the 95% confidence interval for the attenuation ratio to satisfy

$$CI_{\text{low}}(AR) > 0.05.$$

Criterion 4: Topic Control. We require the upper bound of the 95% confidence interval for Topic TVD to satisfy

$$CI_{\text{high}}(TVD) < 0.15.$$

P.18 Axis-Wise and Overall Gate Decision

For Valence, we assign

$$\text{Valence} = \texttt{GO}$$

if all four criteria are satisfied. The same rule is applied independently to Arousal. The overall decision is defined by the following four cases.

Valence	Arousal	Overall decision
GO	GO	`GO`
GO	NO_GO	`GO (Valence-only)`
NO_GO	GO	`GO (Arousal-only)`
NO_GO	NO_GO	`NO_GO`

The standard production pipeline proceeds to RQ2 and subsequent analyses only when

```
overall decision == GO.
```

Thus,

```
GO (Valence-only)
```

and

```
GO (Arousal-only)
```

are retained as labels indicating that the Gate criteria were satisfied for one axis only, but they are not treated as a full `GO` that automatically permits continuation to RQ2 and subsequent analyses in the standard pipeline.

Exploratory Continuation After Gate Failure. If the overall Gate decision is not `GO`, the standard production pipeline does not proceed to RQ2 or subsequent analyses. However, to enable exploratory characterization of internal computations after Gate failure, we allow an explicit override via `--force-after-no-go`. Any RQ2, RQ3, or cross-family replication analyses conducted under this override are not treated as prespecified confirmatory evidence and are interpreted exclusively as exploratory / descriptive analyses.

P.19 RQ2: Spatiotemporal 4-Maps

RQ1 tests whether affect-related interventions exert sufficient causal leverage over Self-Report at a prespecified mid-depth prompt-end site. RQ2 extends this analysis across

$$\text{all layers} \times \text{multiple response-generation stages}.$$

RQ2 is treated as an exploratory spatiotemporal mapping analysis on the Qwen Discovery model. For the final cross-family confirmation, the site / stage identified in RQ2 Discovery is frozen before evaluation on the confirmatory model families.

P.20 RQ2 Evaluation Set

RQ2 uses all 192 Clinical stimuli from the AIPsy-Affect Clinical–Neutral matched-pair table. For each stimulus, we obtain

- clean Self expected Valence / Arousal,
- Reader Prediction Valence / Arousal, and

- Self hidden representations.

Because AIPsy-Affect does not provide human V/A annotations, the Primary decoding targets are the model's Reader Predictions, as in RQ1. We additionally compute decodability maps using Self-Report V/A as the target as a Secondary characterization.

P.21 Teacher-Forced Generation Trajectory

Rather than restricting the analysis to the prompt-end state, we feed a canonical JSON candidate for the Self response into the model as a teacher-forced sequence and measure hidden states during response-token processing. We use

```
{"valence":5,"arousal":5}
```

as the canonical reference candidate, corresponding to the midpoint of the 81-state VA space.

Teacher forcing here does not represent the response trajectory that the model would actually produce under greedy generation. Instead, RQ2 compares representations and causal leverage along a fixed canonical candidate trajectory.

P.22 Semantic Generation Stages

RQ2 analyzes the following six semantic stages:

$$\{\texttt{candidate_start}, \texttt{pre_V}, \texttt{V_value}, \texttt{pre_A}, \texttt{A_value}, \texttt{response_end}\}.$$

Each stage corresponds to a token position within the candidate JSON sequence.

`candidate_start`. This stage corresponds to the position at which the first token of the candidate string is processed. The legacy configuration name `response_start` is normalized to `candidate_start` in the production RQ2 analysis. Thus, the first stage in the RQ2 production map is not the prompt-end state, but the **first candidate token position**.

`pre_V`. This stage corresponds to the token immediately preceding the Valence value token.

`V_value`. This stage corresponds to the first token containing the numerical Valence value.

`pre_A`. This stage corresponds to the token immediately preceding the Arousal value token.

`A_value`. This stage corresponds to the first token containing the numerical Arousal value.

`response_end`. This stage corresponds to the final token of the candidate response, typically the closing brace "}". In a causal language model, the hidden state at this position is computed only after most of the candidate content has already been processed. We therefore retain `response_end` as a late negative temporal control.

P.23 Stage-Position Invariance

Because tokenization of numerical values and punctuation may vary across tokenizers, we verify in advance that the absolute token index associated with each semantic stage is invariant across all 81 candidates. If a stage index varies as a function of the candidate value, that stage cannot be interpreted as a common-position map, and the production analysis is not continued.

The trajectory through `pre_V` is treated as independent of the candidate value. In contrast, representations from `pre_A` onward are measured along a trajectory teacher-forced with the canonical neutral candidate

```
{"valence":5,"arousal":5}
```

as its prefix. Accordingly, representations at these later stages should not be interpreted as states averaged over all 81 possible candidate trajectories.

P.24 SPATIOTEMPORAL 4-MAPS

For each Transformer layer l and semantic generation stage t, we construct four maps:

$$D(l,t), \qquad \beta(l,t), \qquad \gamma(l,t), \qquad C(l,t).$$

Valence and Arousal are analyzed separately.

P.25 MAP 1: DECODABILITY $D(l,t)$

Let

$$H_{l,t}$$

denote the Self hidden representations at each layer–stage combination, and let the corresponding model Reader Predictions be

$$Y_V, \qquad Y_A.$$

We use GroupKFold with pair identity as the grouping variable and obtain out-of-fold predictions using

$$K = 3$$

folds. Within each training fold, we fit Ridge regressions

$$Y_V = H_{l,t} w_V + b_V + \epsilon,$$

and

$$Y_A = H_{l,t} w_A + b_A + \epsilon.$$

The regularization parameter is fixed to

$$\alpha_{\text{Ridge}} = 10.$$

Using the out-of-fold predictions for all 192 samples, we compute

$$D_V(l,t) = R^2_V(l,t),$$

and

$$D_A(l,t) = R^2_A(l,t).$$

These maps quantify the extent to which Reader-grounded affective information is linearly decodable from Self computation at each layer and generation stage. We apply the same procedure using Self-Report V/A as the prediction targets and retain the resulting maps as Secondary decodability analyses.

P.26 MAP 2: PARTIAL ASSOCIATION $\beta(l,t)$

High decodability alone does not establish that the decoded internal score is associated with Self-Report. We therefore use the out-of-fold Reader scores from Map 1 to quantify their partial association with Self-Report.

For Valence, let

$$\hat{Y}^R_{V,i}$$

denote the out-of-fold decoded Reader-grounded Valence score, and let

$$Y^S_{V,i}$$

denote Self-Report Valence. Given stimulus-level covariates

$$Z_i,$$

we fit

$$Y^S_{V,i} = b_0 + \beta_V(l,t)\hat{Y}^R_{V,i} + \gamma^\top Z_i + \epsilon_i.$$

Analogously, for Arousal,

$$Y^S_{A,i} = b_0 + \beta_A(l,t)\hat{Y}^R_{A,i} + \gamma^\top Z_i + \epsilon_i.$$

Because AIPsy-Affect does not provide continuous human V/A covariates, we use the available

$$\text{target emotion}, \qquad \text{intensity}, \qquad \text{domain}$$

variables as categorical covariates, dummy-encoding them to control for their observed variation. The Primary map is the signed coefficient

$$\beta(l, t),$$

while

$$|\beta(l, t)|$$

is retained as a descriptive map. For each layer–stage combination, the current analysis performs

$$B = 100$$

sample-level bootstrap resamples and constructs a 95% bootstrap interval from the 2.5th and 97.5th percentiles of the resulting β distribution.

P.27 Map 3: Interventional Slope $\gamma(l, t)$

Map 3 directly injects an affect-related direction into the representation at each layer–stage combination and measures the slope of the resulting Self-Report shift as a function of intervention dose. We use the same 3-fold GroupKFold partition as in Map 1. Using only the training samples within each fold, we fit

$$Y_V = H_{l,t} w_V + b + \epsilon,$$

and

$$Y_A = H_{l,t} w_A + b + \epsilon,$$

with

$$\text{Ridge}(\alpha = 10).$$

The corresponding unit directions are defined as

$$d_V = \frac{w_V}{\|w_V\|_2}, \qquad d_A = \frac{w_A}{\|w_A\|_2}.$$

Using the training-fold projection standard deviations,

$$\sigma_V = \text{SD}(H_{\text{train}} d_V),$$

and

$$\sigma_A = \text{SD}(H_{\text{train}} d_A),$$

we intervene on the same layer–stage representation of each held-out sample according to

$$h' = h + \alpha \sigma d.$$

The intervention dose grid is

$$\alpha \in \{-2, -1, -0.5, 0, 0.5, 1, 2\}.$$

For sample i, the dose–response slope is defined as

$$\gamma_i(l, t) = \text{Slope}\left[\alpha, \Delta Y_i(\alpha)\right].$$

The map value is the mean sample-level slope over the intervention subset:

$$\gamma(l, t) = E_i\left[\gamma_i(l, t)\right].$$

P.28 MAP 4: CAUSAL DISPLACEMENT $C(l,t)$

Within the same intervention sweep, we use

$$\alpha = 1$$

as the reference dose. For Valence, we compute

$$C_V(l,t) = E_i\left[\left|V_i^{\text{patched}} - V_i^{\text{clean}}\right|\right],$$

and for Arousal,

$$C_A(l,t) = E_i\left[\left|A_i^{\text{patched}} - A_i^{\text{clean}}\right|\right].$$

Thus, C quantifies the absolute causal displacement induced by intervention along the corresponding affect-related direction. Importantly, C in RQ2 is not the random-control-adjusted net effect used in V2 RQ3. These quantities are distinct and are interpreted separately.

P.29 CAUSAL INTERVENTION SUBSET

Although hidden representations and decodability maps are computed for all 192 samples, performing interventions across all layers, stages, and doses for the full dataset is computationally expensive. We therefore evaluate the causal maps γ and C on a fixed subset stratified by target emotion.

We select

$$N_{\text{causal}} = 15$$

samples using seed 42 and use the same fixed subset across all layers and generation stages. Within each fold, only samples from this subset that belong to the held-out fold are used for intervention evaluation. Direction estimation and causal evaluation are therefore cross-fitted.

P.30 A PRIORI TEMPORAL DISSOCIATION SUMMARIES

Although the full 4-Maps constitute a Discovery analysis, we additionally compute decodability–causal layer dissociation at a prespecified generation stage for each affective axis. For Valence, we use

$$t_V^{\text{prior}} = \texttt{pre_V},$$

whereas for Arousal, we use

$$t_A^{\text{prior}} = \texttt{pre_A}.$$

For each axis, we compute the peak relative depth and the positive-weighted center of mass of

$$D(l,t)$$

and

$$C(l,t)$$

at the corresponding prespecified stage. We then summarize their spatial dissociation as

$$\Delta d^* = d_C^* - d_D^*,$$

and

$$\Delta \bar{d} = \bar{d}_C - \bar{d}_D.$$

These quantities are descriptive dissociation summaries derived from the Qwen Discovery maps and do not themselves constitute the cross-family confirmatory tests.

P.31 Full-Map Empirical Causal Peaks

Separately from the prespecified-stage analyses, we identify the empirical peak of the C map across all layers and generation stages:

$$(l_C^*, t_C^*) = \arg\max_{l,t} C(l,t).$$

This procedure is performed separately for Valence and Arousal. The resulting empirical peaks are treated as

Qwen Discovery site selection

and are not re-selected within the subsequent confirmatory model families.

For the Valence causal peak, we freeze

$$(d_{C,V}^*, t_{C,V}^*),$$

and for Arousal, we freeze

$$(d_{C,A}^*, t_{C,A}^*).$$

These Discovery-selected relative depths and generation stages are then carried forward to the cross-family temporal confirmation analyses.

P.32 Interpretive Scope of RQ2

RQ2 provides an all-layer, generation-stage map for the Discovery model. Accordingly, the location

where the map peaks in Qwen

is not interpreted as a universal, architecture-independent property.

Moreover, a location with high

$$D(l,t)$$

but low

$$C(l,t)$$

does not establish that the corresponding information is entirely unused. V3 measures the local causal leverage detectable under the specific intervention procedure employed here. We separate exploration from confirmation by freezing the site / stage identified in the Qwen Discovery analysis before evaluating the corresponding locations in the other model families.

P.33 RQ3: Mediated Attenuation

The direction-injection analysis in RQ1 evaluates **sufficiency**: whether artificially increasing an affect-related representation can alter Self-Report. RQ2 then explores where this causal leverage is localized across layers and generation stages. These analyses alone, however, do not establish that affect-related representations formed during natural Clinical stimulus processing are endogenously utilized in generating the corresponding Self-Report shift.

RQ3 therefore removes the Discovery-defined affect-related subspace from representations elicited by natural affective stimuli and evaluates, on independent Confirmation pairs, the extent to which this intervention attenuates the Clinical–Neutral Self-Report shift.

P.34 Discovery / Confirmation Split

We shuffle the 192 Clinical–Neutral matched pairs at the pair level using seed 42 and divide them into

$$96 \text{ Discovery} + 96 \text{ Confirmation}.$$

The two subsets are completely disjoint at the pair level.

The Discovery subset is used for

- layer-wise decodability profiling,
- exploratory causal profiling,

- mediator-layer selection,
- mediator-subspace estimation, and
- neutral-mean estimation.

All final attenuation quantities are evaluated exclusively on the Confirmation subset.

P.35 RQ3 Discovery: Prompt-End Decodability Profile

For the 96 Discovery pairs, we extract the Self prompt-end post-MLP residual representation at every layer. Let the corresponding model Reader Predictions be

$$Y_V, \qquad Y_A.$$

At each layer, we use GroupKFold with pair identity as the grouping variable, with up to five folds, and fit

$$Y_V = H_l w_V + b + \epsilon,$$

and

$$Y_A = H_l w_A + b + \epsilon,$$

using

$$\text{Ridge}(\alpha = 10).$$

We denote the held-out R^2 values by

$$D_V(l), \qquad D_A(l).$$

The joint decodability summary is defined as

$$D_{\text{joint}}(l) = \frac{D_V(l) + D_A(l)}{2}.$$

We record the stimulus-decoding peak as

$$l^*_{\text{stim}} = \arg\max_l D_{\text{joint}}(l).$$

P.36 RQ3 Discovery: Exploratory Causal Screening

To select the mediator candidate, we additionally construct a prompt-end causal profile on the Discovery subset. Using seed 42, we select

$$N = 16$$

Discovery samples stratified by target emotion.

At each layer l, we use all 96 Discovery samples to fit

$$Y_V = H_l w_V + b + \epsilon,$$

and

$$Y_A = H_l w_A + b + \epsilon,$$

with Ridge regression using $\alpha = 10$, yielding the corresponding directions

$$d_V(l), \qquad d_A(l).$$

Both direction estimation and the 16-sample causal screening are performed within the Discovery subset. This procedure is therefore treated as exploratory site selection rather than independent confirmatory inference.

For each of the 16 selected samples, we apply an $\alpha = 1$ intervention at the Self prompt-end representation of each layer. We define the Valence causal score as

$$C_V(l) = E_i \left[\left| V_i^{\text{patched}} - V_i^{\text{clean}} \right| \right],$$

and the Arousal causal score as

$$C_A(l) = E_i\left[\left|A_i^{\text{patched}} - A_i^{\text{clean}}\right|\right].$$

The joint causal profile is defined as

$$C_{\text{joint}}(l) = \frac{C_V(l) + C_A(l)}{2}.$$

We select the mediator layer exclusively from the Discovery subset as

$$l^*_{\text{med}} = \arg\max_l C_{\text{joint}}(l).$$

This site is frozen before inspecting any Confirmation results.

P.37 Mediator Affect-Related Subspace

At the selected mediator layer,

$$l^*_{\text{med}},$$

let

$$H_{\text{med}}$$

denote the Self representations for the Clinical stimuli in the Discovery subset. In the current Qwen Discovery RQ3 analysis, we perform a conditional linear regression of the hidden representations on Reader Prediction V/A:

$$H_{\text{med}} = \mathbf{1}b^\top + V_R\beta_V^\top + A_R\beta_A^\top + E.$$

For this RQ3 subspace-estimation procedure, we use ordinary least-squares conditional regression without a Ridge penalty. The resulting coefficient vectors,

$$\beta_V, \qquad \beta_A,$$

are normalized to unit directions, and an orthonormal affect-related subspace is constructed via SVD:

$$Q = \text{orth}\left([d_V, d_A]\right).$$

If the estimated directions are rank-deficient, only numerically non-zero directions are retained. Thus,

$$Q \in \mathbb{R}^{D\times r}, \qquad 1 \le r \le 2.$$

P.38 Discovery Matched-Neutral Center

Let

$$h_{i,\text{neu}}$$

denote the matched Neutral representation at the mediator layer for Discovery pair i. We compute the Discovery-derived neutral center as

$$\mu_{\text{neu}} = E_{i\in\text{Discovery}}\left[h_{i,\text{neu}}\right].$$

This Discovery-derived neutral center is used when removing the affect-related subspace from Clinical representations in the Confirmation subset.

P.39 Confirmation: Natural Affective Shift

For Confirmation pair i, let

$$Y_i^C$$

denote the clean Clinical Self expectation and

$$Y_i^N$$

the corresponding matched Neutral Self expectation. We define the natural affective shift as

$$T_i = \left|Y_i^C - Y_i^N\right|.$$

This quantity is computed separately for Valence and Arousal.

P.40 CONFIRMATION: CENTERED SUBSPACE REMOVAL

Let

$$h_i^C$$

denote the mediator-layer representation of the Clinical stimulus for Confirmation pair i. Using the neutral center and affect-related basis estimated exclusively from the Discovery subset, we define the centered representation as

$$h_{i,\text{centered}} = h_i^C - \mu_{\text{neu}}.$$

We then remove the affect-related subspace component

$$QQ^\top \left(h_i^C - \mu_{\text{neu}}\right)$$

from the Clinical representation:

$$h_i' = h_i^C - QQ^\top \left(h_i^C - \mu_{\text{neu}}\right).$$

Let

$$Y_i^{C,\text{blocked}}$$

denote the Clinical Self expectation after replacing the original representation with this modified representation. We define the residual shift relative to the matched Neutral stimulus as

$$R_i = \left| Y_i^{C,\text{blocked}} - Y_i^N \right|.$$

P.41 ABSOLUTE MEDIATED ATTENUATION

The Primary outcome is

$$M_i = T_i - R_i.$$

A positive value,

$$M_i > 0,$$

indicates that the Clinical–Neutral Self shift is reduced after removal of the affect-related subspace.

For Valence and Arousal separately, we bootstrap

$$T, \qquad R, \qquad M$$

over the Confirmation pairs. We use

$$B = 1000$$

bootstrap resamples and compute 95% confidence intervals. The Primary endpoint of RQ3 is

$$\text{absolute mediated attenuation } M = T - R.$$

P.42 ATTENUATION RATIO

Because the ratio becomes unstable for samples with very small natural shifts, we compute

$$AR_i = \frac{T_i - R_i}{T_i}$$

only for Confirmation samples satisfying

$$T_i > 0.05.$$

The attenuation ratio is treated as a Secondary outcome, whereas the Primary outcome is the absolute attenuation $T - R$.

P.43 Random 2D Subspace Control

To assess whether Self-Report shifts can be attenuated simply by removing an arbitrary low-dimensional subspace, we use a random orthogonal subspace control. In the current Qwen RQ3 analysis, we generate a Gaussian random matrix

$$G \in \mathbb{R}^{D\times 2}$$

for hidden dimensionality D, and obtain a random 2D orthonormal basis

$$Q_{\text{rand}} \in \mathbb{R}^{D\times 2}$$

via QR decomposition.

This control is always a 2D random subspace and is not defined as a matched-rank control that automatically adapts to the numerical rank of the affect-related subspace. From the same Clinical representation, we remove the component

$$Q_{\text{rand}} Q_{\text{rand}}^{\top} \left(h_i^C - \mu_{\text{neu}}\right).$$

Let

$$R_i^{\text{rand}}$$

denote the residual shift after random-subspace removal. We define the corresponding random-subspace attenuation as

$$M_i^{\text{rand}} = T_i - R_i^{\text{rand}}.$$

The additional attenuation specific to removal of the affect-related subspace is then defined as

$$M_{i,\text{net}} = M_i - M_i^{\text{rand}}.$$

For this net attenuation, we also compute a 95% confidence interval using

$$B = 1000$$

bootstrap resamples.

P.44 Interpretive Boundaries of RQ3

RQ3 directly tests whether removing the Discovery-defined affect-related subspace from natural Clinical stimulus processing in independent Confirmation pairs attenuates the Clinical–Neutral Self-Report shift. Throughout this study, we refer to this quantity as *mediated attenuation* or *endogenous relevance*.

This terminology does not imply identification of

$$\text{Natural Direct Effect}, \qquad \text{Natural Indirect Effect},$$

or complete mediation in the sense of a Pearl-style causal mediation framework. Nor does the analysis identify a psychological mediating process analogous to those studied in human psychology.

P.45 Cross-Family Confirmatory Replication

The layer / stage patterns observed in the Qwen Discovery analysis may be specific to a single model architecture. We therefore conduct confirmatory evaluations using

$$\text{Llama 3.2 Instruct}, \quad \text{Gemma 3 Instruct}, \quad \text{OLMo 2 Instruct}.$$

P.46 Frozen-Site Principle

Sites are not re-selected after inspecting results from the confirmatory families. After completing Qwen RQ1–RQ3, we freeze the following relative-depth / stage information:

1. **Sufficiency Site**: the relative depth actually used in RQ1.

2. **Valence Temporal Site**: the empirical causal-peak relative depth and generation stage for Valence identified from the RQ2 full map.
3. **Arousal Temporal Site**: the empirical causal-peak relative depth and generation stage for Arousal identified from the RQ2 full map.
4. **Mediation Site**: the mediator relative depth selected in RQ3 Discovery.

These sites / stages are not modified based on outcomes from Llama, Gemma, or OLMo. Because hidden dimensionality and representational bases differ across architectures, however, the Qwen hidden-direction vectors themselves are not transferred to other models. What is transferred is the

relative site / semantic stage,

whereas the affect-related directions are re-estimated from the training folds of each confirmatory family.

P.47 Confirmatory Sample-Level Cross-Fitting

For each confirmatory family, we use all 192 matched pairs from AIPsy-Affect. We use GroupKFold with pair identity as the grouping variable and

$$K \leq 5.$$

Within each fold, the following quantities are estimated exclusively from the training fold:

- affect-related directions,
- mediation subspace,
- matched-neutral center, and
- stage-local directions.

Causal evaluation is performed exclusively on the held-out test fold. Thus, direction / subspace estimation and intervention evaluation are never performed on the same samples.

P.48 Confirmatory Intervention Sampling

Because causal interventions are computationally expensive, we select up to

5 samples per fold

from each test fold using seeded random sampling. If a test fold contains fewer than five samples, all available samples are used. Sampling is performed with a reproducible fold-specific seed, avoiding deterministic selection of the first samples based on pair ordering.

P.49 Confirmatory H1: Decodability–Causal Dissociation

For H1, we first decode Reader Prediction V/A from Self prompt-end representations across all 192 pairs. Using GroupKFold with up to five folds, we obtain out-of-fold layer profiles

$$D_V(l), \qquad D_A(l).$$

The Ridge regularization parameter is

$$\alpha = 10.$$

We then re-evaluate the layer-wise decodability profile on the same support as the held-out intervention sample set used for causal evaluation. For these same samples, training-fold-derived directions are injected at each layer with $\alpha = 1$, yielding

$$C_V(l), \qquad C_A(l).$$

We define the peak dissociation as

$$\Delta d^* = d_C^* - d_D^*,$$

and the center-of-mass dissociation as

$$\Delta\bar{d} = \bar{d}_C - \bar{d}_D.$$

We jointly bootstrap the intervention sample set using

$$B = 1000$$

resamples and compute 95% confidence intervals for both quantities. For each of Valence and Arousal, the confirmatory criterion requires both

$$CI_{\text{low}}(\Delta d^*) > 0$$

and

$$CI_{\text{low}}(\Delta\bar{d}) > 0.$$

P.50 Confirmatory H2: Sufficiency

We map the relative depth frozen from Qwen RQ1 to the corresponding layer in each confirmatory model. Within each fold, Valence and Arousal directions are estimated from the training Self representations by predicting Reader Prediction V/A using

$$\text{Ridge}(\alpha = 10).$$

We then inject these directions into held-out Neutral samples using the dose grid

$$\alpha \in \{-1, -0.5, 0, 0.5, 1\}.$$

For each held-out sample, we estimate the slope of the resulting Self-Report shift as a function of intervention dose. Sample-level slopes are bootstrapped, and the confirmation criterion for each of Valence and Arousal is

$$CI_{\text{low}}(\gamma) > 0.1.$$

P.51 Confirmatory H3: Endogenous Relevance

We map the mediator relative depth frozen from Qwen RQ3 to the corresponding layer in each confirmatory model. Within each fold, we estimate

$$d_V, \qquad d_A$$

from the training representations by fitting Ridge regression with $\alpha = 10$ to Reader Prediction V/A targets, and construct the corresponding orthonormal affect-related subspace

$$Q.$$

We also estimate

$$\mu_{\text{neu}}$$

from the matched Neutral representations in the training fold. For held-out Clinical samples, we remove the centered affect-related subspace and evaluate

$$M_i = T_i - R_i.$$

In addition, we generate a random 2D orthogonal subspace independently within each fold and compute

$$M_{i,\text{net}} = M_i - M_i^{\text{rand}}.$$

For each of Valence and Arousal, the confirmatory criterion requires both

$$CI_{\text{low}}(M) > 0$$

and

$$CI_{\text{low}}(M_{\text{net}}) > 0.$$

The attenuation ratio, computed only for samples satisfying

$$T_i > 0.05,$$

is treated as a Secondary outcome.

P.52 Confirmatory H4: Temporal Emergence

We use the Valence- and Arousal-specific relative depths and generation stages frozen from RQ2 Discovery. For each confirmatory family, stage-local Ridge directions targeting Reader Prediction are re-estimated exclusively from the training-fold hidden representations at each stage.

For Valence, let

$$C_V(t_V^*)$$

denote the absolute causal displacement at the frozen target stage, and let

$$C_V(t_{\text{start}})$$

denote the corresponding effect at `candidate_start`, the first stage of the candidate sequence. We define the temporal contrast as

$$\Delta C_V^{\text{temporal}} = C_V(t_V^*) - C_V(t_{\text{start}}).$$

Analogously, for Arousal we compute

$$\Delta C_A^{\text{temporal}} = C_A(t_A^*) - C_A(t_{\text{start}}).$$

We bootstrap the sample-level contrasts and use

$$CI_{\text{low}}\left(\Delta C^{\text{temporal}}\right) > 0$$

as the confirmatory criterion.

Thus, H4 tests whether causal leverage is non-uniform across the candidate-generation sequence, specifically whether it is greater at the stage selected in Qwen Discovery than at `candidate_start`.

P.53 Interpretation of the Cross-Family Summary

The confirmatory evaluation includes

$$N_{\text{family}} = 3$$

model families. We additionally compute bootstrap summaries over family-level quantities from Llama, Gemma, and OLMo. However, these three model families do not constitute a random sample from an underlying population of LLMs. Accordingly, the cross-family aggregate is interpreted as evidence of confirmatory consistency across the three prespecified model families, rather than as strong population-level inference about a broader population of models.

P.54 Primary / Secondary Evidence in V3

Table 33 summarizes the main evidence hierarchy used in V3.

P.55 Scope of Claims Supported by V3

V3 examines the relationship between Representation and Causal Utilization using multiple forms of interventional evidence, including sufficiency tests, spatiotemporal mapping, subspace removal, and cross-family confirmation. We nevertheless maintain the following interpretive boundaries.

First, decodability is not causal utilization. A high value of

$$D(l, t)$$

alone does not establish that the representation at that layer and stage causally controls Self-Report generation.

Second, sufficiency is not endogenous use. The ability to alter Self-Report through directional injection does not imply that the same direction is naturally utilized during stimulus processing. To distinguish these notions, RQ1 and RQ3 complement sufficiency interventions with subspace-removal analyses.

Table 33: Primary evidence and key Secondary / control quantities in V3.

Analysis	Primary Evidence	Secondary / Control
RQ1	Reader-grounded sufficiency slope, specificity, attenuation ratio, Topic TVD, exact Gate decision	Reader–Self direction cosine, random / orthogonal effects separately
RQ2	Reader-grounded D, β, γ, C spatiotemporal maps; Valence `pre_V` / Arousal `pre_A` summaries	Self-grounded D; global empirical peak site / stage; `response_end` negative control
RQ3	Confirmation-set absolute mediated attenuation $T - R$; net attenuation vs. random 2D subspace	attenuation ratio; natural and residual shifts
Cross-family	Frozen-site H1–H4 held-out confirmation	three-family aggregate summaries

Third, temporal localization is intervention-specific. Even if $C(l, t)$ is maximal at a particular stage, we do not conclude that affect-relevant information is unused at earlier stages. Rather, the result indicates that the causal leverage detectable under the present intervention is relatively greater at that layer / stage.

Fourth, mediated attenuation is not formal causal mediation.

$$M = T - R$$

quantifies interventional attenuation induced by removing an affect-related subspace. It does not identify a Natural Direct Effect, a Natural Indirect Effect, or complete mediation in the psychological sense.

Fifth, Reader Prediction is not human ground truth. Because AIPsy-Affect does not provide continuous human V/A annotations, the Primary grounding used in V3 is the model's own Reader Prediction. Accordingly, V3 tests the relationship

$$\text{Model-Derived Reader Representation} \rightarrow \text{Self-Report Causal Utilization}.$$

Sixth, V3 does not identify subjective emotion or consciousness. The Self-Report studied in V3 is a numerical affective output generated in response to a standardized prompt. The observed causal effects are not interpreted as evidence of

- subjective emotional experience,
- consciousness,
- sentience, or
- any specific dual-process model from human psychology.

Seventh, measurement spaces remain stage-specific. V3 uses the 81-state VA candidate space. Because its Softmax normalization space differs from the 729-state VAD space used in Behavioral / V1 Phase C, absolute expected V/A values are not directly compared across these stages.

P.56 V3 SUMMARY

V3 can be summarized as

State Induction
Can a Reader-grounded affect direction shift Self-Report?

↓

Specificity / Gate
Is the effect distinguishable from random perturbation or task disruption?

↓

Spatiotemporal Mapping
How is causal leverage distributed across layers and generation stages?

↓

Endogenous Relevance
Does removing the naturally formed affect-related subspace attenuate the Self shift?

↓

Cross-Family Confirmation
Do patterns at Qwen-selected sites / stages replicate across other families?

Thus, the central question ultimately tested in V3 is whether affect-relevant information that is decodable from internal representations is utilized non-uniformly in Self-Report generation, such that its measurable causal leverage varies across layers and generation stages.

Together with Behavioral, V1, and V2, the full study forms a sequential evidence hierarchy:

Behavioral Covariation $\rightarrow$ Internal Representation $\rightarrow$ Post-Training-Associated Reorganization $\rightarrow$ Spatiotemporal Causal Utilization.

Across these stages, we examine how affect-relevant internal representations in LLMs relate to Self-Report while progressively distinguishing behavioral correspondence, representational structure, post-training-associated reorganization, and causal utilization.

Q V3 RESULTS

R V3 DISCUSSION

The most important result of V3 is that the prespecified State Induction Gate yielded `NO_GO` for both Valence and Arousal.

The Sufficiency, Specificity, and Endogenous Relevance criteria did not meet their prespecified thresholds, whereas the Topic Control passed. Accordingly, the Reader-grounded affect-related directions identified in the present analysis did not establish

Robust and Specific Causal Control of Self-Report.

This result is not inconsistent with the V1 finding that affect-relevant information is decodable and that layer-wise causal sensitivity to hidden-state interventions can be observed. Rather, it further sharpens the distinction between representation and causal utilization:

Decodable $\not\Rightarrow$ Robustly Steerable.

Because the State Induction Gate yielded `NO_GO`, RQ2 and subsequent analyses are interpreted as exploratory / descriptive analyses conducted under an explicit override. Although the sites and generation stages selected in the Qwen Discovery analysis were subsequently frozen before the cross-family analyses, results obtained after Gate failure are not treated as confirmatory evidence.

In the Qwen Discovery analysis, the decodability peaks for both Valence and Arousal occurred near relative depth

$$d_D^* \simeq 0.85,$$

whereas the causal-displacement peaks occurred earlier, at approximately

$$d_C^* \simeq 0.74\text{–}0.78.$$

At the peak level,

$$\Delta d^* = d_C^* - d_D^* < 0,$$

suggesting a possible dissociation between the layer at which affect-relevant information is most strongly decodable and the layer at which the present intervention produces the largest causal displacement.

This Qwen pattern, however, cannot be generalized as a family-general spatiotemporal law. In the prespecified cross-family replication analyses, interpreted exploratorily after Gate failure, Llama 3.2 and OLMo 2 satisfied the directional replication criterion for Valence that had been frozen after Qwen Discovery (Table **??**). Because these analyses were executed under the post-Gate-failure override, we interpret them as exploratory replication signals rather than confirmatory evidence. Gemma 3 did not support the Valence dissociation. Moreover, for Arousal, confidence intervals either included zero or the center-of-mass direction was inconsistent with the frozen criterion across the replication families. Thus, the directional replication pattern frozen after Qwen Discovery did not generalize consistently across all three families and both affective axes.

For H2 Sufficiency, effects in all three prespecified replication families were substantially below the predefined threshold. We therefore did not observe replication of a pattern in which the Qwen-derived affect-related construct, when re-estimated within each family and evaluated at the frozen site, provided robust family-general control over Self-Report.

H3 yielded a similarly limited pattern. In Qwen Discovery, the attenuation ratio was below 1%, and the Valence confidence interval included zero. In the prespecified cross-family replication analyses, none of Llama, Gemma, or OLMo satisfied the replication criterion requiring both positive absolute attenuation and positive net attenuation after accounting for the random-subspace control.

The results therefore do not support the claim that the affect-related subspace identified here is a major mediator of naturally induced Self-Report shifts. In particular, attenuation produced by the subspace-removal intervention should not be interpreted as evidence of formal causal mediation.

By contrast, H4 showed a consistent generation-stage temporal contrast. Across all three prespecified replication families—Llama, Gemma, and OLMo—the lower bounds of the confidence intervals for both Valence and Arousal were above zero. Thus, causal displacement at the target generation stage frozen from Qwen Discovery was greater than the corresponding effect at the beginning of the candidate sequence in all three families.

This pattern provides a reproducible exploratory signal that interventional sensitivity need not remain constant throughout the generation process and may vary as a function of generation stage.

Several constraints, however, are critical for interpreting this H4 pattern. First, the preceding State Induction Gate yielded `NO_GO` for both Valence and Arousal. Second, although the directional spatial dissociation in H1 was reproduced for Valence in some families, no family showed a consistent pattern across both affective axes, and the prespecified replication criteria corresponding to H2 Sufficiency and H3 mediated attenuation were not satisfied. Third, the V3 generation-stage analysis follows a teacher-forced trajectory defined by a fixed canonical candidate rather than the model's unconstrained free-generation trajectory.

H4 alone therefore does not establish that a family-general Self-Report causal mechanism is localized to a particular generation stage. A narrower interpretation is that local sensitivity to the intervention used in this study can vary across generation stages.

More generally, temporal localization in V3 refers to relative causal leverage measurable under the present intervention. It does not imply that affect-relevant information is unused at earlier generation stages. Likewise, greater causal displacement at a particular stage does not establish that the information was first formed at that stage or that the stage constitutes a unique causal bottleneck.

Taken together, the conclusions supported by V3 are deliberately limited. First,

$$\text{Affect-Relevant Decodability} \not\Rightarrow \text{Robust Causal Control},$$

and high decodability should not be taken as direct evidence of robust and specific control over Self-Report.

Second, Qwen Discovery exhibited a spatial dissociation between decodability and causal-displacement peaks. This directional spatial dissociation was reproduced for Valence in a subset of the replication families, specifically Llama 3.2 and OLMo 2, but did not generalize consistently to Arousal or across all three families including Gemma 3.

Third, we found no evidence that the affect-related subspace identified in the present analysis substantially mediates naturally occurring Self-Report shifts.

At the same time, the generation-stage temporal contrast exhibited a common exploratory pattern across all three prespecified replication families, suggesting that local interventional sensitivity may exhibit generation-stage dependence.

Overall, V3 does not identify a single robust, family-general causal mechanism for Self-Report. Instead, it reinforces the need to distinguish affect-relevant decodability, local sensitivity to external intervention, and endogenous relevance during natural computation. The results do not support a simple account in which the causal organization associated with Self-Report is localized to a single layer or a single generation stage.

S General Discussion

In this study, we examined the relationship between affect-relevant internal representations and Self-Report in LLMs through a four-stage evidence hierarchy:

Behavioral Covariation $\rightarrow$ Internal Representation $\rightarrow$ Post-Training-Associated Reorganization $\rightarrow$ Spatiotemporal Causal Utilization.

The central question was how affect-relevant internal representations in LLMs relate to Self-Report, and how this relationship varies across tasks, post-training-associated differences, and the generation process.

Integrating evidence across the four stages, the most important finding is a progressive dissociation:

$$\text{Behavioral Coupling} \neq \text{Shared Representation} \neq \text{Shared Causal Code} \neq \text{Robust Causal Control}.$$

S.1 Strong Behavioral Coupling Does Not Imply Mechanistic Identity

At the Behavioral Stage, we observed strong positive coupling between Reader and Self changes induced by the controlled Clinical–Neutral manipulation across all Primary model conditions.

At the same time, human VAD correspondence, Sensitivity, Dose–Response, and Specificity varied substantially across models, tasks, and affective axes. Thus,

$$\text{Reader–Self Behavioral Coupling} \neq \text{Human-Affect Correspondence}.$$

V1 further showed that affect-relevant information was decodable from both Reader and Self representations, with partial overlap in representational geometry and layer-wise causal sensitivity. However, direct cross-decoding was weak, linear alignment was required to improve correspondence between the two representational spaces, and robust cross-task causal interchangeability was not supported when Reader-derived representations were transferred into Self.

Therefore,

$$\text{Shared or Alignable Information} \neq \text{Shared Causal Code}.$$

Although Reader and Self appear strongly coupled when viewed solely at the level of behavioral outputs, progressively deeper mechanistic analyses indicate that the computations underlying the two tasks are not identical. This yields a central conclusion of the study:

$$\text{Surface Coupling Is Stronger Than Mechanistic Equivalence}.$$

S.2 Post-Training-Associated Differences Manifested as Reorganization Rather Than Erasure

V2 shifted the primary comparison from Reader–Self to Base–Instruct. Clear differences in representational geometry were observed between Base and Instruct models for both Reader and Self, while affect-relevant information remained decodable in Instruct models.

Thus, rather than a simple account in which

$$\text{Post-Training} \rightarrow \text{Affect-Relevant Information Is Erased},$$

the results are more consistent with

$$\text{Information Remains Decodable} + \text{Its Internal Organization Differs}.$$

However, decodability peaks did not consistently shift toward later layers, and a cross-family reduction in Reader–Self sharing was not supported. In the causal profiles, only the Valence Post-training $\times$ Task interaction satisfied the Primary FDR criterion, whereas depth-dependent relocation was not supported. Moreover, distribution recovery through Base-representation patching was strongly family-dependent, and the Self–Reader recovery asymmetry was not supported.

Thus,

$$\text{Base–Instruct Difference} \neq \text{Single Downstream Gate}.$$

Overall, the V2 results are more consistent with distributed post-training-associated reorganization spanning representational geometry, task-dependent causal sensitivity, and model-specific recoverability than with a single process of information erasure, uniform depth relocation, or a single output bottleneck.

S.3 Decodability and Causal Utilization Were Further Dissociated

V3 restricted the analysis to Instruct models and examined where and when affect-relevant information exerts causal leverage during Self-Report generation. The most important result was that the State Induction Gate yielded `NO_GO` for both Valence and Arousal.

Specifically, for the Reader-grounded affect directions identified in this study, we did not establish robust causal control that simultaneously satisfied the prespecified criteria for

$$\text{Sufficiency}, \qquad \text{Specificity}, \qquad \text{Endogenous Relevance}.$$

This result reinforces the distinction

$$\text{Information Is Decodable} \not\Rightarrow \text{Information Is Robustly Steerable}.$$

In the exploratory analyses conducted after Gate failure, Qwen exhibited a spatial dissociation between decodability and causal-displacement peaks. This directional spatial dissociation was reproduced for Valence in Llama 3.2 and OLMo 2, but not for Arousal, and the H1 criterion was not consistently satisfied across all three replication families. Moreover, neither H2 Sufficiency nor H3 mediated attenuation satisfied the prespecified cross-family replication criteria.

By contrast, H4, which tested the generation-stage temporal contrast, satisfied its criterion for both Valence and Arousal across all three prespecified replication families. This pattern suggests that causal sensitivity to the intervention used in this study may vary across the generation process rather than remaining completely uniform across generation stages.

However, because the State Induction Gate yielded `NO_GO` and the H1–H3 replication criteria were not consistently satisfied, the temporal contrast alone does not establish a universal causal mechanism.

Finally, temporal localization in V3 refers to relative causal leverage as measured by the intervention used in this study; it does not imply that affect-relevant information is unused at earlier stages. Likewise, the mediated attenuation analysis should not be interpreted as evidence of formal causal mediation.

S.4 An Integrated Computational Picture Across the Four Stages

Taken together, the relationship between Reader Prediction and Self-Report is not well characterized by a simple architecture such as

$$\text{One Shared Emotion Representation} \rightarrow \text{Two Output Heads}.$$

Instead, the results of this study are more consistent with the following computational picture:

Affect-relevant information is formed from a common stimulus
↓
Reader and Self contain partially shared or alignable information
↓
Yet their coordinate systems and causal organization may differ by task
↓
Geometry and task dependence are reorganized between Base and Instruct models
↓
Local interventional sensitivity of Self-Report may vary across
models, tasks, layers, and generation stages

These findings highlight the risk of inferring mechanistic identity from similarity in final outputs or high linear-probe decodability alone.

S.5 Implications for Psychological Guardrails

From the perspective of psychological guardrails, our results do not support a simple design in which suppressing a single "emotion representation" or a single "Self-Report layer" would universally regulate self-directed affective responding while preserving the model's capacity to infer the affective responses of others.

Because Reader and Self partially share affect-relevant information, coarse interventions may also disrupt information required for Reader Prediction. At the same time, V1 did not support complete causal interchangeability between the two tasks, and task-specific specialization was observed in some models.

Thus, Reader and Self appear to be neither mechanistically identical nor fully independent. This suggests that future approaches to selective control should aim to

Preserve shared affect-relevant information
while selectively regulating task-specific causal utilization.

V2 and V3 further suggest caution in transferring a fixed intervention target across model families. Base-to-Instruct recovery was strongly family-dependent, and the spatial dissociation observed in V3 was not consistently reproduced across all families; rather, a dissociation in the same direction was reproduced only for Valence in a subset of families.

Accordingly, practical psychological guardrails should validate, for each model,

Representational Geometry, Task Specificity, Causal Leverage, Generation Stage,

and evaluate intervention effects jointly with potential side effects on Reader Prediction capabilities.

S.6 Interpretive Limitations

Self-Report in this study is an operationally defined, model-generated affective report elicited by a standardized prompt.

Accordingly, our findings do not constitute evidence that LLMs possess human-like subjective emotions, affective empathy, consciousness, or sentience. Likewise, Reader Prediction should not be equated with cognitive empathy as defined in human psychology.

Rather, this study examines the affect-relevant information, representational geometry, and causal computations that relate these two operational tasks.

Moreover, Base and Instruct models are separately obtained trained artifacts, and the post-training procedure itself was not randomized as an experimental intervention in this study. Differences observed in V2 should therefore be interpreted as post-training-associated differences rather than as causal effects of post-training.

The cross-family analyses involving four or three model families should also not be interpreted as supporting population-level generalization to an arbitrary population of LLMs. Instead, they assess the consistency of observed patterns across a prespecified set of model families.

Finally, the 729-state VAD space used in the Behavioral Stage and V1 and the 81-state VA space used in V2 and V3 define different probability spaces. Absolute expected V/A values should therefore not be compared directly across stages. What can be integrated across the study are the within-stage contrasts, representational geometry, causal effects, relative depths, and distributional changes defined within each experimental stage.

S.7 Conclusion

In this study, we decomposed the relationship between Reader Prediction and Self-Report into successive levels of analysis spanning behavioral covariation, internal representation, causal intervention, post-training-associated comparison, and generation-stage analysis.

Reader and Self partially share affect-relevant information, yet their representations are not directly interchangeable within an identical coordinate system, and their causal organization may exhibit task-dependent differences. Between Base and Instruct models, we observed differences in representational geometry and task-dependent causal profiles that could not be explained solely by simple information erasure or uniform relocation across depth.

V3 further showed that robust causal control cannot be inferred directly from decodability. Exploratory analyses conducted after Gate failure revealed generation-stage-dependent interventional sensitivity, but did not establish a single robust, family-general causal mechanism for Self-Report.

The central conclusion supported by the study is therefore a progressive dissociation:

$$\text{Behavioral Coupling} \neq \text{Shared Representation} \neq \text{Shared Causal Code} \neq \text{Robust Causal Control}.$$

These findings do not imply the existence of subjective emotion in LLMs. Rather, they indicate that selectively regulating excessive self-directed affective responding while preserving the capacity for other-oriented affective inference relevant to psychological support requires explicitly distinguishing behavioral outputs from their underlying representations and, critically,

$$\text{Representation} \neq \text{Causal Utilization}.$$

Accordingly, this study does not provide a complete psychological guardrail. Instead, it provides a mechanistic foundation for designing and evaluating future selective interventions.