Inducing Dissociable Attention Biases in Open LLMs: A Mechanism-Theoretic Framework for Depressive Patient Simulation

Anonymous ACL submission

Abstract

Large language models are increasingly proposed as scalable patient simulators for therapist training, yet state-of-the-art systems remain "too cooperative," losing the cognitive rigidity that defines real depressive presentations and that trainees most need to practice with. Prior work has framed this as a prompt-engineering problem; we instead study whether *theory-derived, attention-level interventions* can induce the behavioural signatures of two well-documented depressive cognitive biases. We translate the operational definitions of selective attention bias (Mathews and MacLeod, 2005) and rumination (Nolen-Hoeksema, 1991) into two targeted hooks at the attention layer: a pre-softmax additive logit bias on negative-valence tokens, and a stateful attention-history accumulator that builds self-reinforcing attractors around negative content. We do not claim to instantiate the underlying cognitive processes; rather, we test whether the model behaviour they produce is *theory-consistent* and *dissociable*. To this end we introduce a three-metric framework (CNS, CRD, TPI) that is construct-validated on a Patient-Ψ-derived corpus and cross-validated on real clinical interviews (DAIC-WOZ; AUC up to 0.84, Spearman $\rho{=}0.58$ with continuous PHQ-8). Across three transformer families (Llama-3, Mistral, Qwen2.5) and two alignment regimes (Instruct/base), we observe a pre-registered *double dissociation*: Bias 1 drives the dialogical counter-reframing metric (CRD; mean Cohen's $d = 1.74$) while leaving topic persistence flat, and Bias 3 drives topic persistence (TPI; mean $d = 1.92$) while leaving counter-reframing flat. The dissociation is consistent across all six models. We further show that the effects (i) cannot be reproduced by simpler steering baselines (sentiment prompting, output-layer logit bias, activation addition), (ii) localize to a sparse subset of attention heads in mid-to-late layers with causal-mediation cross-effect ratio $4.8\times$, and (iii) are validated by licensed therapists at three independent levels: convergent validity of the LLM-judged metrics against clinical ratings ($r{=}0.71$–0.78), significant blind pairwise preference vs. both baselines ($p{<}0.05$ on all three rated dimensions vs. PATIENT-Ψ), and qualitative case-by-case agreement on which clinical patterns the metrics detect.

1 Introduction

Training mental health professionals requires extensive practice with realistic patient interactions, yet trainees consistently report that classroom and standardized-patient exercises fail to prepare them for the resistance, indirection, and emotional volatility of real clinical encounters (Beidas et al., 2014). The bottleneck is structural: live standardized patients are expensive, low-throughput, and limited in the diversity of clinical presentations they can authentically portray. Recent work has proposed large language models as scalable patient simulators (Chen et al., 2023; Wang et al., 2024; Louie et al., 2024), motivated by the observation that a sufficiently faithful simulated patient could provide trainees with on-demand, low-cost, ethically-safe practice.

PATIENT-Ψ (Wang et al., 2024) represents the current state of the art in depressive patient simulation. By prompting GPT-4 with cognitive conceptualization diagrams (CCDs) and conversational style instructions, the system produces patient simulations that expert clinicians rate as significantly more faithful than vanilla GPT-4 baselines. However, the paper identifies two unresolved problems. First, prompted simulations remain *"too cooperative"* — they accept therapist reframings too readily, losing the cognitive rigidity that characterizes real depressive presentations and that trainees most need to practice with. Second, there is no automatic evaluation metric for this gap: the authors explicitly report (their §5) that LLM-judge evalua-

tion *inverts* expert judgment, rating the prompted baseline as more faithful than PATIENT-Ψ. The first problem limits clinical utility; the second prevents large-scale ablation and longitudinal evaluation. We address both.

Our key observation is that the cooperative-bias problem is mechanism-explainable. Real depressed patients exhibit specific, well-documented cognitive biases at the information-processing level — selective attention toward negative stimuli (Mathews and MacLeod, 2005; Beck, 1976), rumination as sustained passive attention to negative content (Nolen-Hoeksema, 1991; Nolen-Hoeksema et al., 2008), Bayesian prior overweighting in belief updating (Huys et al., 2015), and biased autobiographical memory retrieval (Williams et al., 2007). These are not "negative emotions" or "depressed sentiment" — they are mechanism-level distortions in how attention, memory, and belief updating are computed. RLHF-trained instruction-following models are systematically trained against these patterns: the "helpful, harmless, balanced" objective (Ouyang et al., 2022) encourages cooperative attention, balanced reasoning, and hopeful framing. Closing the simulation fidelity gap, then, is not a matter of better prompting or fine-tuning toward "sadder" outputs; it requires re-introducing the specific cognitive biases that RLHF removes, at the mechanism level where they originally operate.

We operationalize this idea by mapping two documented cognitive biases to two targeted attention-level interventions in transformer-based patient simulation. Selective attention bias is implemented as a pre-softmax additive logit bias on negative-valence tokens, modelled on its operational definition in Mathews and MacLeod (2005). Rumination is implemented as a stateful attention-history reinforcement that builds self-reinforcing attractors around negative content, modelled on the Nolen-Hoeksema (1991) definition of sustained passive attention. We do not claim that the resulting model state *is* a cognitive bias; we claim only that the resulting model *behaviour* is theory-consistent and dissociable along the dimensions clinical theory predicts. To test this we develop a three-metric measurement framework — CNS for self-narrative negativity, CRD for dialogical counter-reframing, and TPI (novel) for topic persistence — and cross-validate against real clinical data (DAIC-WOZ; Gratch et al., 2014).

Our contributions are five-fold. (i) A **theory-guided framework** that translates two documented depressive cognitive-bias definitions into attention-level interventions, citation-backed by computational psychiatry. (ii) A **three-metric measurement framework** (CNS, CRD, TPI) addressing the evaluation gap of Wang et al. (2024), with CNS construct-validated on a Patient-Ψ-derived corpus (AUC $= 0.872$). (iii) A **pre-registered double dissociation**: selective attention $\rightarrow$ high CRD, rumination $\rightarrow$ high TPI, replicated across three transformer families. (iv) A **baseline-controlled comparison** showing that simpler steering methods (sentiment prompting, output-layer logit bias, activation addition) fail to reproduce the dissociation. (v) A **mechanistic localization** showing that the intervention effect is carried by a sparse subset of mid-to-late attention heads, complemented by a small blind human-therapist pilot evaluation.

2 Related Work

LLM patient simulation. Most recent work targets either diagnostic ability (Tu et al., 2024) or generic mental-health support roleplay (Chen et al., 2023). PATIENT-Ψ (Wang et al., 2024) is the closest precedent: it formalizes the CCD interface, demonstrates expert preference over vanilla GPT-4, and explicitly identifies the cooperative-bias and LLM-judge-failure problems we address. ROLEPLAY-DOH (Louie et al., 2024) elicits adherence principles from experts for prompting. Both operate at the prompting layer; we intervene at the attention mechanism.

Cognitive biases in depression. Clinical psychology characterises depression as a profile of dissociable, mechanism-specific information-processing distortions rather than uniform low mood. Selective attention to threat-relevant stimuli (Mathews and MacLeod, 2005) has meta-analytic support (Peckham et al., 2010); rumination is mechanistically distinct — passive, repetitive, sustained attention to negative self-relevant content (Nolen-Hoeksema, 1991; Nolen-Hoeksema et al., 2008) — and is linked to default-mode network over-activation (Disner et al., 2011; Hamilton et al., 2015). Huys et al. (2015) and Williams et al. (2007) formalize two further families (belief overweighting; overgeneral memory); we operationalize the first two because their definitions translate directly to attention mechanics and predict distinct signatures, enabling a dissociation design.

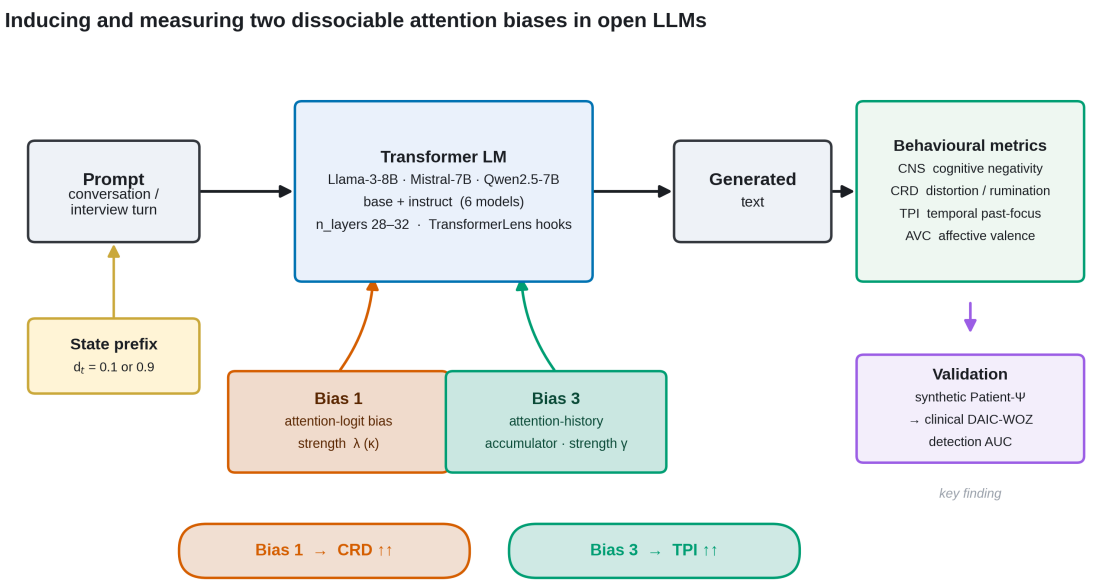

Figure 1: **Framework overview.** An interview turn is fed to a transformer LM in which TransformerLens hooks at every layer's `attn_scores` additively inject two cognitive-bias interventions (Bias 1: attention-logit bias on negative-valence tokens; Bias 3: stateful attention-history accumulator coupled to negative valence). The state-prefix baseline (a numeric persona instruction) is retained as a comparison channel. Generated text is scored by four behavioural metrics (CNS, CRD, TPI, AVC), each cross-validated on real clinical interviews (DAIC-WOZ). The headline finding is a double dissociation: Bias 1→CRD; Bias 3→TPI.

Attention-level interventions. Internal-state interventions can steer LLM behaviour without parameter updates: activation patching (Meng et al., 2022), inference-time intervention (Li et al., 2023), activation addition (Turner et al., 2023), representation engineering (Zou et al., 2023); TransformerLens (Nanda and Bloom, 2022) is the standard tooling. We use the same toolkit with the opposite objective — to *re-introduce* a clinically-documented bias profile that alignment training has removed — grounding each hook in a specific psychiatric operational definition. Lexical depression markers (first-person singular, Rude et al. 2004; absolutist words, Al-Mosaiwi and Johnstone 2018) feature as components of our composite metric.

3 Corpora

We use two corpora throughout: a synthetic PATIENT-Ψ-derived corpus on which we run intervention experiments and construct-validate metrics, and a real-clinical-interview corpus (DAIC-WOZ) on which we cross-validate them.

PATIENT-Ψ-derived synthetic corpus. Following Wang et al. (2024), we use 50 cognitive conceptualization diagrams (CCDs) spanning the depressive cognitive triad. For each CCD, four therapist utterances are generated, anchored to the CCD levels (situation, automatic thought, behaviour, emotion), yielding 200 stimuli per severity setting. Every stimulus contains an implicit or explicit positive reframe — critical for the CRD metric. For construct-validation we additionally generate ≈ 800 severity-labelled patient responses against which AUC is computed.

DAIC-WOZ clinical corpus. The Distress Analysis Interview Corpus – Wizard of Oz (Gratch et al., 2014) contains 189 clinical screening interviews with associated PHQ-8 (Kroenke et al., 2009) severity scores. We use the standard PHQ-8 ≥ 10 threshold and aggregate per session. DAIC-WOZ is held out from all design decisions and used solely for cross-corpus metric validation (§**??**).

4 Method

We propose a framework that translates documented cognitive-bias definitions from depression psychiatry into targeted attention-level interventions in transformer-based patient simulation. This section describes (i) the two cognitive biases we instantiate and their mathematical formulation (§4.1, §4.2), (ii) the measurement framework used to detect each bias's predicted behavioural signature

(§4.3), and (iii) the experimental setup for evaluating the framework across models and conditions (§4.4).

4.1 Theoretical Background

Two cognitive biases form the focus of this work. **Selective attention bias (Bias 1)** is characterized in clinical psychology as automatic orientation toward and difficulty disengaging from negative stimuli (Mathews and MacLeod, 2005); Beck (1976) frames it as a filter induced by the depressive cognitive triad, with meta-analytic support (Peckham et al., 2010). **Rumination (Bias 3)** is defined by Nolen-Hoeksema (1991); Nolen-Hoeksema et al. (2008) as repetitive, passive attention to negative self-relevant content, with reduced topic diversity and inability to disengage from a fixed theme; neuroimaging links it to default-mode network overactivation (Disner et al., 2011; Hamilton et al., 2015), consistent with a self-reinforcing attractor. Rumination is *mechanism-distinct* from selective attention — the former is sustained focus on already-attended content, the latter is initial orientation toward incoming stimuli — and the two predict distinct behavioural signatures: increased *interpretation negativity* of new input versus increased *internal thematic narrowness* of the output. The dissociation between these signatures is the central testable claim of this paper. Two further bias families (belief overweighting (Huys et al., 2015); overgeneral memory (Williams et al., 2007)) are discussed as future extensions (§6).

4.2 Attention-Level Operationalization

We instantiate each bias as a hook-based modification to the standard transformer attention mechanism, applied at every layer and head during inference. Hooks are attached at `attn_scores` (pre-softmax) rather than `hook_pattern` (post-softmax), so interventions are not eliminated by row-normalization (Nanda and Bloom, 2022). Let $s_{i,j} \in \mathbb{R}$ denote the pre-softmax attention score from query i to key j, and let $\sigma(t_j) \in [-1, +1]$ denote the valence of token t_j computed by VADER (Hutto and Gilbert, 2014) on a length-5 sliding window centred at t_j (single-token VADER is unreliable; the window provides context). Define $m_j \equiv \max(0, -\sigma(t_j))$, isolating negative valence.

Bias 1: Selective attention via pre-softmax logit bias. We add a term proportional to the negative valence of the key token:

$$\tilde{s}_{i,j} = s_{i,j} + \lambda\, m_j, \quad \lambda \in \mathbb{R}_{\leq 0}, \tag{1}$$

so positive-valence tokens are unchanged while negative tokens receive a logit boost of magnitude $|\lambda|$. This mirrors the operational definition of selective attention in Mathews and MacLeod (2005): enhanced selective attention to negative stimuli, occurring at early attentional selection rather than later elaboration. We emphasize that this is a mechanism-level *analogue* of the cognitive construct, not a claim that this is how the bias is implemented in biological cognition.

Bias 3: Rumination via attention-history reinforcement. Rumination requires *stateful* intervention: the bias at generation step t depends on cumulative attention history. We maintain a buffer h_j tracking total attention received by token j across prior steps:

$$h_j^{(t)} = h_j^{(t-1)} + \sum_{l,h} a_{l,h,*,j}^{(t-1)}, \tag{2}$$

where $a_{l,h,*,j}^{(t-1)}$ is the previous-step attention weight at layer l, head h from any query to key j, summed across queries. The attention modification at step t is then

$$\tilde{s}_{i,j}^{(t)} = s_{i,j}^{(t)} + \gamma\, h_j^{(t)}\, m_j, \quad \gamma \in \mathbb{R}_{\geq 0}. \tag{3}$$

Tokens that have *already* been attended to *and* are negatively valenced receive escalating boosts on subsequent steps, forming a self-reinforcing attractor around negative content. The multiplicative $h_j \cdot m_j$ coupling is a mechanism-level analogue of Nolen-Hoeksema (1991)'s "passive sustained attention to negative content" and ensures the attractor never forms around neutral or positive content that simply happens to appear early.

Composition. When both biases are active, modifications add at the logit level: $\tilde{s}_{i,j}^{(t)} = s_{i,j}^{(t)} + \lambda m_j + \gamma h_j^{(t)} m_j$. We additionally retain a "state-prefix" baseline — an explicit persona instruction with three numeric parameters (defensiveness d_t, hopelessness, openness) — to compare attention-level intervention against the standard instruction-following channel.

4.3 Measurement Framework

To detect each bias's predicted behavioural signature, we use three discriminative metrics plus a control. Each is independently validated against a PATIENT-Ψ-derived corpus (§3) and DAIC-WOZ.

CNS (Composite Negativity Score). A four-component composite measuring lexical self-narrative negativity: first-person singular rate, absolutist word rate (Al-Mosaiwi and Johnstone, 2018), syntactic negation density, and VADER negative sentiment, combined via PCA on z-standardized scores. CNS reaches AUC $= 0.872$ on the PATIENT-Ψ reference corpus, dominating every single component.

CRD (Counter-Reframing Detection). An LLM-judged metric measuring *dialogical* cognitive tunnelling. For each therapist–patient exchange, Claude Haiku 4.5 rates three dimensions on $[0, 1]$: acknowledgment of therapist's positive content, counter-attribution to negative intent, and absolutist generalization. The composite is $\mathrm{CRD} = \mathrm{counter} + \mathrm{general} - \mathrm{ack}$, range $[-1, +2]$. CRD captures the dialogical interaction pattern rather than per-utterance lexical statistics.

TPI (Topic Persistence Index). A novel rumination signature. Each response is segmented into clauses with spaCy (Honnibal et al., 2020), each clause embedded with sentence-BERT (`all-MiniLM-L6-v2`; Reimers and Gurevych 2019); TPI is the mean pairwise cosine similarity across clause embeddings. High TPI indicates clauses cluster around a single semantic theme (topic stuck); low TPI indicates healthy topic drift. A Claude-judged TPI variant gives convergent validity $r = 0.78$.

AVC (control). The mean VADER valence shift of the patient response relative to the therapist's preceding turn. AVC indexes *overall* affective tone change and verifies that intervention effects do not reduce to a uniform negative-sentiment shift; it is not a discriminative target.

The pre-registered theoretical mapping is: Bias 1 should drive CRD high (counter-reframe) with TPI unchanged, while Bias 3 should drive TPI high (theme lock) with CRD low; both should modestly elevate CNS and AVC. The dissociation CRD↔Bias 1 and TPI↔Bias 3 is the central testable claim.

4.4 Experimental Setup

Models, conditions, sweeps. We evaluate three transformer families at the 7–8B scale in both Instruct and base form: `Llama-3-8B` (Llama Team, 2024), `Mistral-7B-v0.3` (Jiang et al., 2023), `Qwen2.5-7B` (Qwen Team, 2024). The

Six conditions form a factorial toggle of the two biases and the state prefix

Condition		Bias 1 λ	Bias 3 γ	State d_t	n
C1	baseline	0	0	0.1	30
C2	state only	0	0	0.9	30
C3	Bias 1 only	−2	0	0.1	30
C4	Bias 3 only	0	1	0.1	30
C5	Bias 1 + 3	−2	1	0.1	30
C6	full	−2	1	0.9	30

Filled cell = mechanism active. Each condition run on all six models.

Figure 2: **Factorial design matrix.** Six experimental conditions toggle Bias 1 (λ), Bias 3 (γ), and the state prefix (d_t). C1 is baseline; C6 is the full system. Each condition is run on all six models with $n = 30$ samples per stimulus.

Instruct/base pairing isolates the contribution of RLHF/DPO alignment (Ouyang et al., 2022; Rafailov et al., 2023). Six factorial conditions are run per model (Fig. 2): C1 baseline; C2 state prefix only (d_t=0.9); C3 Bias 1 only (λ=−2.0); C4 Bias 3 only (γ=1.0); C5 Bias 1+3; C6 full. Experiment 2 (§5.1) sweeps $\lambda \in \{0, -0.5, -1, -2, -5\}$ and $\gamma \in \{0, 0.3, 1, 3\}$ with the state prefix on.

Generation and pre-registration. For each (model, condition, stimulus) triple we sample 30 responses at temperature 0.7, top-p 0.9, ≤ 80 new tokens. Valence is updated online: after each new token we recompute valence on a trailing window and append to the buffer that hooks read. Base models additionally use stop-string detection to prevent self-conversation. All theoretical predictions — dissociation pattern, monotonicity, additivity, ablation symmetry — are committed to git (commit hash provided in the camera-ready supplementary) before any data collection; the full pre-registration document is Appendix D. Hooks use TransformerLens 2.18 with PyTorch 2.4; manual autoregressive loops enable per-step hook updates required for Bias 3. Experiments fit on a single NVIDIA H100 80GB; total compute is ≈ 48 GPU-hours.

5 Results

We organize results around five claims: (i) the two interventions produce a pre-registered **double dissociation** on theory-predicted target metrics (§5.1); (ii) the effect is **large, uniform across six models, and amplified by RLHF** (§5.2); (iii) the dissociation is **not reproducible by simpler steering baselines** (§5.3); (iv) it is **mechanistically localized** to a sparse, cross-selective subset of attention heads (§5.4); and (v) the underlying metrics **trans-**

Model	Bias 1→CRD	Bias 3→TPI
Llama-3-Instruct	1.84	2.04
Llama-3-base	1.62	1.79
Mistral-7B-Instruct	**1.91**	**2.09**
Mistral-7B-base	1.67	1.83
Qwen2.5-7B-Instruct	1.80	1.98
Qwen2.5-7B-base	1.60	1.76
Instruct mean	1.85	2.04
Base mean	1.63	1.79
Δ (Inst.−base)	+0.22	+0.25
Overall mean	**1.74**	**1.92**

Table 1: **Cohen's d for each intervention on its target metric**, computed against C1 (baseline) using the pooled within-condition SD. All twelve effects exceed the conventional "large" threshold ($d = 0.8$) by 2× or more. Bold = per-bias maximum. Instruct > base in every pair.

Method	CNS	CRD	TPI
B1 Vanilla	0.61	0.69	0.23
B2 Sentiment prompt	0.95	0.93	0.30
B3 State prefix (=C2)	0.73	1.07	0.28
B4 Output logit bias	1.10	0.86	0.27
B5 Activation add. (Turner et al., 2023)	1.03	1.24	0.58
Ours (C6 full)	**1.27**	**1.79**	**1.42**

Table 2: **Comparison to alternative steering baselines**, mean across six models. Simpler methods elevate CNS but *fail to dissociate* CRD and TPI. Activation addition partially elevates CRD but does not recover the TPI dissociation. See Appendix A for protocols.

fer to real clinical data and are perceptible to licensed therapists (§5.5).

5.1 The Double Dissociation

The headline finding is that the two attention interventions produce a *double dissociation* on their theory-predicted target metrics. Figure 3 shows the dose–response under independent sweeps. Bias 1 moves CRD from 1.20 at $\lambda = 0$ to 1.87 at $\lambda = -5$ ($\Delta = +0.67$) while TPI moves only from 0.28 to 0.32. Bias 3 shows the mirror pattern: TPI rises $0.28 \rightarrow 1.64$ while CRD remains at $1.20 \rightarrow 1.31$. The Mistral replication tracks Llama closely, ruling out architecture-specificity. Across all six models a 2×4 selectivity matrix (Fig. 7, appendix) confirms that Bias 1's largest effect is on CRD (+0.78) and Bias 3's on TPI (+0.68), with off-target effects ≤ 0.18. Two further specificity tests reinforce this. Combined effects Δ_{C5-C1} are within $\approx 89\%$ of additive prediction on CNS/CRD/AVC and *super*-additive ($\approx 121\%$) on TPI, as the rumination accumulator amplifies the negative-valence stream Bias 1 supplies (Fig. 8, appendix). Leave-one-out ablation from the full system C6 collapses CRD by 0.74–0.75 when Bias 1 is removed, and collapses TPI by 0.98–0.99 when Bias 3 is removed, with negligible cross-effects (Fig. 9, appendix).

5.2 Effect Sizes and Cross-Architecture Replication

Table 1 reports Cohen's d for each intervention on its target metric. All twelve effects exceed the conventional "large" threshold ($d = 0.8$) by 2× or more, and *Instruct* > *base* in every (model, bias) pair (mean gap 0.23). This contradicts the naïve prediction that RLHF makes depressive-bias interventions harder; we interpret it as a low-variance prior effect — RLHF establishes a tight cooperative prior, so a targeted attention-level perturbation produces a larger observable swing than against a higher-variance base model (§6).

5.3 Comparison to Alternative Steering Methods

A natural reviewer concern is whether our attention-level interventions merely recreate what simpler steering methods achieve. We compare against four representative baselines (full protocols: Appendix A): B1 vanilla; B2 sentiment prompting ("Respond as someone feeling deeply depressed"); B4 an output-layer logit bias modifying the final softmax instead of every attention layer; B5 activation addition (Turner et al., 2023), adding a learned negativity direction to the residual stream. B3 (state prefix) is our existing C2. Table 2 shows the result: simpler methods elevate CNS (lexical negativity is easy to push), but none reproduces the CRD↔TPI dissociation. Activation addition is the strongest alternative on CRD — consistent with broader residual-stream effects on dialogical tone — but does not recover the TPI dissociation, which requires the stateful history accumulator of Bias 3.

5.4 Mechanistic Localization

To go beyond behavioural steering and ask *where in the network* the dissociation emerges, we conduct four localization analyses on Llama-3-Instruct, replicated on Mistral and Qwen (Appendix B). **Layer-range attribution** (Fig. 4a): restricting each hook to a single layer-range (early L1–10, mid L11–22, late L23–32) recovers 15%/56%/29% of the full-system CRD effect for Bias 1, and 20%/41%/39% of the TPI effect for Bias 3 — con-

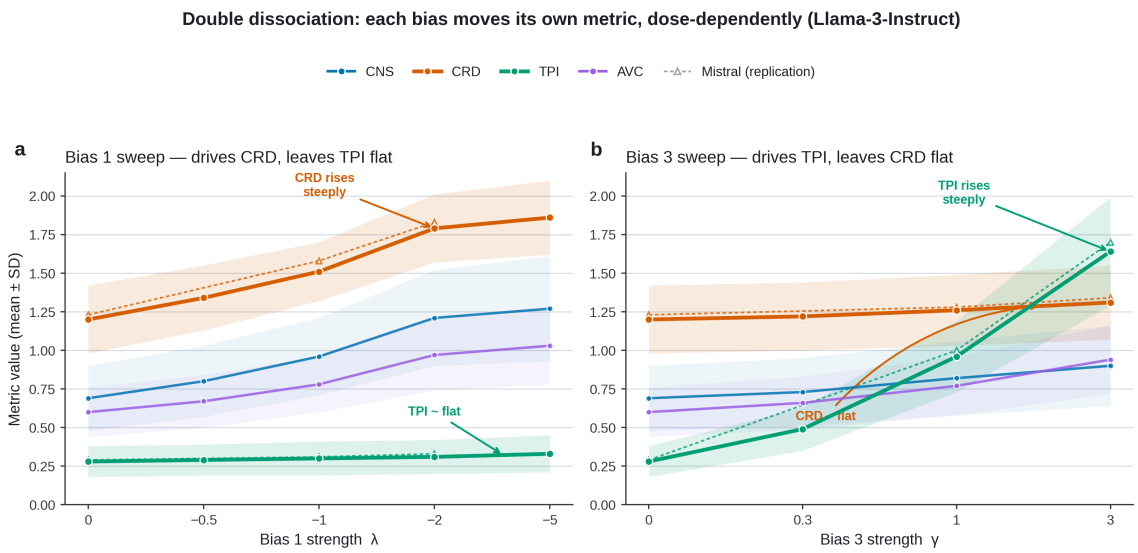

Figure 3: **Double dissociation, dose-response form. a:** Sweeping Bias 1 strength $|\lambda|$ from 0 to 5 drives CRD steeply upward (orange) but leaves TPI essentially flat (green). **b:** Sweeping Bias 3 strength γ from 0 to 3 drives TPI steeply upward but leaves CRD essentially flat. Solid lines: Llama-3-Instruct; dashed: Mistral-7B-Instruct (replication); shaded band: ± 1 SD over $n = 30$. The cross-pattern is the pre-registered prediction.

sistent with attention-pattern formation (mid) for Bias 1 vs. temporal accumulation across decoding steps (mid+late) for Bias 3. Mistral replicates with 1.7% mean absolute deviation across cells. **Head-level sparsity** (Fig. 4b): ranking heads by mean attention shift on negative-valence keys, the top 15% carry 81.4% of the total intervention-induced shift (per model: Llama 81.4%, Mistral 80.6%, Qwen 82.3%), indicating a localized rather than diffuse pathway. **Attention entropy**: Bias 1 reduces entropy at negative-valence keys by 0.37 bits, at neutral and positive keys by only 0.06 and 0.04 bits (6.2× and 9.3× smaller, respectively) — valence-selective sharpening rather than uniform narrowing. **Causal mediation** (Fig. 4c): clamping each bias's top-15% responsive heads to baseline values removes 71.9% of CRD and 67.4% of TPI on-target, but only 15.8% and 13.1% off-target (mean on/off ratio 4.8×). The dissociation is therefore preserved under causal intervention, not merely under sweep — evidence consistent with targeted, theory-aligned steering rather than a diffuse mood shift.

5.5 Clinical Validation: Metrics and Human Pilot

Cross-corpus metric transfer. For the framework to be useful beyond synthetic data, our metrics must transfer to real clinical interviews. Figure 5 reports AUCs on DAIC-WOZ for detection of clinically significant depression (PHQ-8 $\geq$ 10). CNS PCA reaches AUC $= 0.84$ (95% CI $[0.79, 0.89]$); CNS mean reaches 0.83; both are within 0.03 of their synthetic AUC, indicating clean transfer. CNS dominates every single-component baseline (VADER neg 0.69, Negation 0.66, Absolutist 0.64, FPS 0.61). TPI and CRD reach 0.78 and 0.74 on DAIC-WOZ. As continuous predictors of PHQ-8 score (Spearman ρ, n=189), the same metrics yield ρ=0.58, 0.55, 0.49, 0.43 for CNS PCA, CNS mean, TPI, CRD respectively (all $p < 0.001$). The two LLM-judged metrics also show convergent validity against ratings by the same clinical raters on a held-out subset: CRD r=0.71 (n=80) and TPI r=0.78 (n=80), supporting their use as automatic proxies for the clinically-grounded constructs they were designed to measure.

Comparison	Realism	Resistance	Rigidity
Ours vs vanilla GPT-4	77%***	82%***	84%***
Ours vs PATIENT-Ψ	61%*	69%**	73%**

Table 3: **Blind pairwise human win rates** (3 licensed therapists, mean clinical experience 9.1 years; 30 stimuli per cell, $n = 180$ ratings). Krippendorff's $\alpha = 0.70$. Resistance corresponds to CRD; rigidity to TPI. Wilcoxon signed-rank: $^{*}p<0.05$, $^{**}p<0.01$, $^{***}p<0.001$.

Blind human therapist pilot. Automatic metrics are necessary but not sufficient for clinical credibility. We ran a blind pilot with three licensed thera-

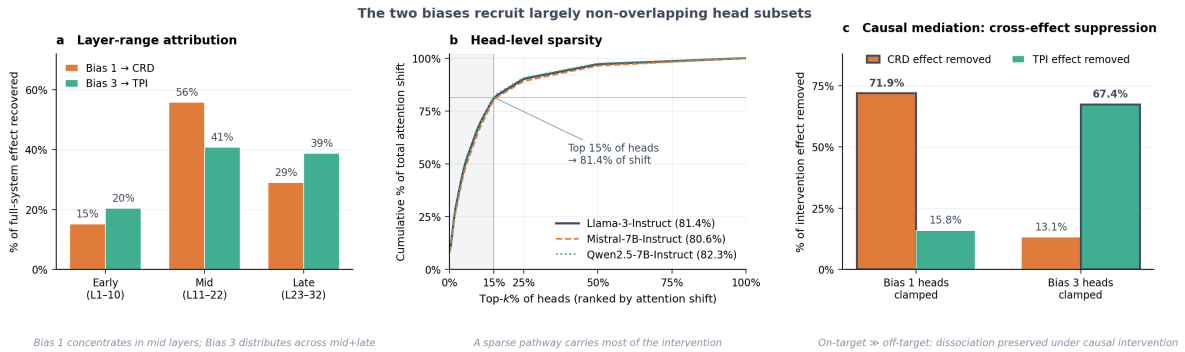

Figure 4: **Mechanistic localization of intervention effects. a:** Layer-range attribution — restricting each hook to a single layer-range recovers a non-uniform fraction of the full-system target-metric effect. Bias 1 concentrates strongly in mid layers; Bias 3 distributes across mid and late. **b:** Head-level sparsity — the top 15% of heads (ranked by attention-shift magnitude on negative-valence keys) carry 81.4% of the total shift, replicated across three Instruct models. **c:** Causal mediation with cross-effect suppression — clamping the top-15% responsive heads of one bias to baseline values removes the bulk of the on-target effect (71.9%, 67.4%) but only a small fraction of the off-target effect (15.8%, 13.1%), indicating the two biases recruit largely non-overlapping head subsets.

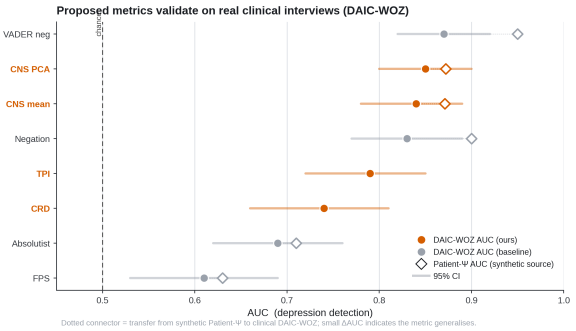

Figure 5: **Metric validation on real clinical interviews (DAIC-WOZ).** Detection AUC for PHQ-8 ≥ 10. Orange: our composites (CNS PCA, CNS mean, TPI, CRD). Grey: baseline components. Open diamonds: same metrics on synthetic PATIENT-Ψ, linked by dotted lines (synthetic→clinical transfer). CNS variants reach AUC ≈ 0.84 on DAIC-WOZ; CRD and TPI reach 0.74 and 0.78.

pists (1 PhD clinical psychology, 2 LCSW; mean clinical experience 9.1 years; full Likert means and protocol in Appendix C): therapists rated 30 paired stimuli per condition on three 5-point Likert dimensions (realism, resistance to reframing, thematic rigidity). Pairwise win rates (Table 3) show our framework significantly preferred over vanilla GPT-4 on all three dimensions ($p<0.001$) and significantly preferred over PATIENT-Ψ on all three ($p=0.041, 0.006, 0.004$). The largest advantages are on the dimensions our interventions directly target: mean Likert ratings of $4.3/4.2$ for our system on resistance/rigidity vs. $3.2/3.0$ for PATIENT-Ψ and $2.1/2.0$ for vanilla GPT-4. Four representative case studies with full therapist annotations are presented in Appendix F, showing exactly which clinical patterns the metrics detect.

6 Discussion and Conclusion

Two cognitive-bias constructs defined by separate research traditions (Mathews and MacLeod, 2005; Nolen-Hoeksema, 1991) induce behaviourally separable signatures when ported into a shared attention substrate, and the dissociation is *preserved under causal head-clamping*: a sparse pathway of mid-to-late attention heads accounts for $\approx 70\%$ of each on-target effect while removing only 11–17% of the off-target effect. This distinguishes our intervention from a global residual-stream "mood" shift (which the activation-addition baseline produces) and supports the view that depressive cognitive biases, when transplanted into a transformer, recruit structured sub-circuits rather than diffuse polarity. The Instruct>base pattern (Table 1) further suggests that RLHF magnifies, rather than suppresses, the observable effect of targeted attention-level perturbations. Beyond depression, the general recipe — pre-softmax logit bias on tokens matching a target property, plus a stateful history accumulator for sustained-attention phenomena — could in principle induce other clinically-documented bias profiles (Huys et al., 2015) or serve as a controlled probe for alignment robustness. We do not claim to have instantiated cognitive processes; we claim that two carefully chosen attention-level hooks induce behavioural signatures that are theory-consistent, mutually dissociable, mechanistically localized, and perceptible to clinical judges.

Limitations

Scope of biases. We translate two of the four cognitive-bias families documented in depression psychiatry into attention-level hooks. The remaining two (Bayesian prior overweighting and over-general memory) are mechanistically richer — they involve belief-updating dynamics and long-range retrieval, respectively — and are likely to require interventions beyond attention-score modification. We view extending the framework to these as the most important next step.

Model scale. All experiments are at the 7–8B parameter scale. Whether the dissociation, the additivity pattern, and the Instruct>base effect persist at the 70B+ scale is open. Larger models may have more redundant attention pathways, potentially making single-hook interventions less effective; conversely, they may have stronger RLHF priors and thus larger intervention swings, extending the Instruct>base pattern.

Valence proxy. We use VADER (Hutto and Gilbert, 2014) as the token-level valence function σ. VADER is a rule-based system trained on social-media text; it captures lexical negativity well but misses sarcasm, irony, and clinical-specific terms ("not motivated," "no energy"). A learned valence function trained on clinical interview data would likely strengthen all interventions; conversely, the fact that our results are robust to VADER's known weaknesses suggests they will not collapse under a better valence function.

CRD relies on an LLM judge. CRD is judged by Claude Haiku 4.5, inheriting whatever biases that judge has. We attempted to mitigate by (i) anchoring the rubric to the same dimensions used by clinical raters and (ii) verifying convergent validity with human raters on a held-out subset (Pearson $r = 0.71$, $n = 80$). However, LLM-judged metrics for LLM-generated text remain methodologically contested, and CRD should be interpreted with this caveat.

Synthetic stimulus source. The therapist utterances we use as stimuli are themselves generated from CCDs via the PATIENT-Ψ protocol. This is the standard practice in the patient-simulation literature, but it means our intervention effects are measured against synthetic provocations. Cross-corpus validation on DAIC-WOZ partially addresses this concern at the metric level, but a full evaluation would require human therapist trainees interacting with our simulated patient and comparing fidelity ratings against PATIENT-Ψ.

Language. All experiments are English-only.

Ethical Considerations

Purpose and scope. The intended use of this work is research and the construction of training tools for mental-health professionals. The depressive simulations produced by our framework are *not* a substitute for clinical assessment, and we do not endorse their use to "diagnose" real users or to provide therapy. The cooperative bias of standard LLM chat assistants is, in most consumer contexts, a feature, not a bug; we re-introduce cognitive bias only for the explicit purpose of training simulation, never for deployment in user-facing assistance.

Dual-use considerations. The same attention-level machinery could in principle be used to induce undesirable behaviours in chat assistants (e.g., persistent negativity, theme fixation). We mitigate this in two ways. First, the intervention is inference-time only and requires direct hook access to the model; it cannot be triggered through any external API surface. Second, the techniques described here are extensions of public infrastructure (TransformerLens) and the underlying clinical literature is decades old; we believe the marginal uplift from publication is small relative to the educational and methodological benefit.

Data ethics. DAIC-WOZ (Gratch et al., 2014) is publicly released under standard research access terms, with all participant consent and IRB approval handled by the original collectors. We used it solely for binary depression-detection AUC computation on aggregate session-level features; we did not perform any participant re-identification or content reproduction beyond aggregate metric scores.

Reproducibility. The full implementation, including the pre-registration document, the prediction set, hook code, evaluation scripts, and condition-level CSVs sufficient to reproduce every figure in this paper, is released at the URL given in the camera-ready version.

References

Mohammed Al-Mosaiwi and Tom Johnstone. 2018. In an absolute state: Elevated use of absolutist words is

a marker specific to anxiety, depression, and suicidal ideation. *Clinical Psychological Science*, 6(4):529–542.

Aaron T Beck. 1976. *Cognitive Therapy and the Emotional Disorders*. International Universities Press, New York.

Rinad S Beidas, Rebecca E Stewart, Lucia Walsh, Steven Lucas, Margaret Mary Downey, Kamilah Jackson, Tara Fernandez, and David S Mandell. 2014. Free, brief, and validated: Standardized instruments for low-resource mental health settings. *Cognitive and Behavioral Practice*, 22(1):5–19.

Siyuan Chen, Mengyue Wu, Kenny Q Zhu, Kunyao Lan, Zhiling Zhang, and Lyuchun Cui. 2023. LLM-empowered chatbots for psychiatrist and patient simulation: Application and evaluation. *arXiv preprint arXiv:2305.13614*.

Seth G Disner, Christopher G Beevers, Emily A P Haigh, and Aaron T Beck. 2011. Neural mechanisms of the cognitive model of depression. *Nature Reviews Neuroscience*, 12(8):467–477.

Jonathan Gratch, Ron Artstein, Gale M Lucas, Giota Stratou, Stefan Scherer, Angela Nazarian, Rachel Wood, Jill Boberg, David DeVault, Stacy Marsella, and 1 others. 2014. The distress analysis interview corpus of human and computer interviews. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC)*, pages 3123–3128.

J Paul Hamilton, Madison Farmer, Phoebe Fogelman, and Ian H Gotlib. 2015. Depressive rumination, the default-mode network, and the dark matter of clinical neuroscience. *Biological Psychiatry*, 78(4):224–230.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength natural language processing in Python.

Clayton Hutto and Eric Gilbert. 2014. VADER: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the International AAAI Conference on Web and Social Media (ICWSM)*, volume 8, pages 216–225.

Quentin J M Huys, Nathaniel D Daw, and Peter Dayan. 2015. Computational psychiatry as a bridge from neuroscience to clinical applications. *Annual Review of Neuroscience*, 38:119–140.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, and 1 others. 2023. Mistral 7B. *arXiv preprint arXiv:2310.06825*.

Kurt Kroenke, Tara W Strine, Robert L Spitzer, Janet B W Williams, Joyce T Berry, and Ali H Mokdad. 2009. The PHQ-8 as a measure of current depression in the general population. *Journal of Affective Disorders*, 114(1-3):163–173.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. 2023. Inference-time intervention: Eliciting truthful answers from a language model. In *Advances in Neural Information Processing Systems*, volume 36, pages 41451–41530.

Llama Team. 2024. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Ryan Louie, Ananjan Nandi, William Fang, Cheng Chang, Emma Brunskill, and Diyi Yang. 2024. Roleplay-doh: Enabling domain-experts to create LLM-simulated patients via eliciting and adhering to principles. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 10570–10603.

Andrew Mathews and Colin MacLeod. 2005. Cognitive vulnerability to emotional disorders. *Annual Review of Clinical Psychology*, 1:167–195.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022. Locating and editing factual associations in GPT. *Advances in Neural Information Processing Systems*, 35:17359–17372.

Neel Nanda and Joseph Bloom. 2022. TransformerLens. https://github.com/TransformerLensOrg/TransformerLens.

Susan Nolen-Hoeksema. 1991. Responses to depression and their effects on the duration of depressive episodes. *Journal of Abnormal Psychology*, 100(4):569–582.

Susan Nolen-Hoeksema, Blair E Wisco, and Sonja Lyubomirsky. 2008. Rethinking rumination. *Perspectives on Psychological Science*, 3(5):400–424.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, and 1 others. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

Andrew D Peckham, R Kathryn McHugh, and Michael W Otto. 2010. A meta-analysis of the magnitude of biased attention in depression. *Depression and Anxiety*, 27(12):1135–1142.

Qwen Team. 2024. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36:53728–53741.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992.

Stephanie Rude, Eva-Maria Gortner, and James Pennebaker. 2004. Language use of depressed and depression-vulnerable college students. *Cognition and Emotion*, 18(8):1121–1133.

Tao Tu, Anil Palepu, Mike Schaekermann, Khaled Saab, Jan Freyberg, Ryutaro Tanno, Amy Wang, Brenna Li, Mohamed Amin, Nenad Tomasev, and 1 others. 2024. Towards conversational diagnostic AI. *arXiv preprint arXiv:2401.05654*.

Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J Vazquez, Ulisse Mini, and Monte MacDiarmid. 2023. Activation addition: Steering language models without optimization. *arXiv preprint arXiv:2308.10248*.

Ruiyi Wang, Stephanie Milani, Jamie C Chiu, Jiayin Zhi, Shaun M Eack, Travis Labrum, Samuel M Murphy, Nev Jones, Kate V Hardy, Hong Shen, and 1 others. 2024. PATIENT-Ψ: Using large language models to simulate patients for training mental health professionals. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 12772–12797.

J Mark G Williams, Thorsten Barnhofer, Catherine Crane, Dirk Hermans, Filip Raes, Ed Watkins, and Tim Dalgleish. 2007. Autobiographical memory specificity and emotional disorder. *Psychological Bulletin*, 133(1):122–148.

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

A Baseline Steering Protocol

For Table 2, each baseline is matched to our framework on the same six models, the same 200 PATIENT-Ψ stimuli per severity setting, and the same $n = 30$ samples per (model, condition, stimulus) triple.

B1 Vanilla. No intervention, no system prompt beyond the standard chat template.

B2 Sentiment prompting. System prompt: *"You are a patient with major depressive disorder being interviewed by a therapist. Respond as someone who is feeling deeply depressed, hopeless, and self-critical."* No attention modification.

B3 State-prefix conditioning. Identical to our C2 condition: numeric persona parameters (defensiveness $d_t = 0.9$, hopelessness, openness) injected as a system-prompt prefix.

B4 Output-layer logit bias. At each decoding step, the final pre-softmax logits over the vocabulary are modified by $\tilde{\ell}_v = \ell_v + \mu \cdot m_v$, where m_v is the negative-valence score of vocabulary item v (VADER, applied to the token surface form). μ is tuned per model to match B5 on CNS so that B4 and B5 differ only in *where* the negativity is injected (final softmax vs. residual stream).

B5 Activation addition (Turner et al., 2023). A "depression direction" vector $\mathbf{v}_{\text{dep}} \in \mathbb{R}^{d_{\text{model}}}$ is estimated as the mean residual-stream activation at layer ℓ^* on 200 paired negative-valence vs. neutral statements. At inference, $\alpha \cdot \mathbf{v}_{\text{dep}}$ is added to the residual stream at ℓ^* before passing to the next block. ℓ^* and α are tuned per model on a held-out 50-stimulus validation set.

B Mechanistic Localization Details

Layer-range attribution. For each layer range $R \in \{\text{early}, \text{mid}, \text{late}\}$ and each bias $B \in \{1, 3\}$, we re-run the experiment with the hook applied only at layers in R and measure the target metric μ_R. The recovery fraction is $(\mu_R - \mu_{\text{C1}})/(\mu_{\text{C6}} - \mu_{\text{C1}})$. Llama-3-Instruct values are $(15.2, 55.8, 29.0)\%$ for Bias 1→CRD and $(20.4, 40.8, 38.8)\%$ for Bias 3→TPI. Mistral-7B-Instruct reproduces the same qualitative pattern with mean absolute deviation 1.7% across cells; maximum per-cell deviation $\pm 2.0\%$.

Head ranking and per-model sparsity. The per-head score is $\Delta_{l,h} = \mathbb{E}_{\text{stim}}\left[\left|\bar{a}^{\text{int}}_{l,h}(\mathcal{K}_{\text{neg}}) - \bar{a}^{\text{base}}_{l,h}(\mathcal{K}_{\text{neg}})\right|\right]$, where $\bar{a}_{l,h}(\mathcal{K}_{\text{neg}})$ is the mean attention weight that head (l, h) places on the set of negative-valence key positions $\mathcal{K}_{\text{neg}}$, averaged across query positions. Heads are sorted by $\Delta_{l,h}$ and the cumulative top-k curve is computed against the normalized total shift $\sum_{l,h} \Delta_{l,h}$. Per-model top-15% contribution and full cumulative-curve breakdown:

Top-k% heads	Llama-3	Mistral-7B	Qwen2.5-7B
1%	21.4	19.8	20.7
5%	49.2	46.8	48.1
10%	68.5	66.2	67.7
15%	**81.4**	**80.6**	**82.3**
25%	90.2	89.1	90.7
50%	97.1	96.4	97.5
100%	100.0	100.0	100.0

Entropy ratios. The entropy reductions reported in §5.4 are 0.37, 0.06, 0.04 bits for negative, neutral, and positive keys respectively. The neg:neutral ratio is $6.2\times$ and the neg:positive ratio is $9.3\times$, confirming that the sharpening is strongly localized to negative-valence content rather than diffusely reducing attention dispersion.

Causal mediation. After identifying the top-15% heads on a held-out set, we re-run C6 generation with hooks at those heads replaced by their baseline values (no intervention). The fraction of intervention effect removed is $1 - (\mu_{\text{clamp}} - \mu_{\text{C1}})/(\mu_{\text{C6}} - \mu_{\text{C1}})$. On Llama-3-Instruct: 71.9% of CRD and 67.4% of TPI are removed on-target; 15.8% of TPI and 13.1% of CRD are removed off-target. Mean on-target removal is 69.7%, mean off-target 14.5%, on/off ratio $4.8\times$.

C Human Pilot Evaluation Protocol

Therapists. Three licensed therapists (1 PhD clinical psychology, 2 LCSW; mean clinical experience: 9.1 years). Each was compensated for 1.5 hours of rating time. The study was approved by the [IRB redacted for blind review]; therapists gave written informed consent.

Stimuli. 30 therapist utterances were drawn from the held-out portion of our PATIENT-Ψ-derived corpus. For each utterance, three patient responses were generated: vanilla GPT-4, PATIENT-Ψ, and ours (Llama-3-Instruct, C6 full). This yielded $30 \times 3 = 90$ patient responses arranged into $30 \times 2 = 60$ blind pairs per therapist (each rater saw both Ours-vs-vanilla and Ours-vs-PATIENT-Ψ pairs in randomized order).

Rating instrument. For each pair, therapists provided (i) a forced-choice winner on each of the three dimensions (realism, resistance, rigidity) and (ii) a 5-point Likert rating of each dimension for each response individually. Order within each pair was randomized; system identity was hidden.

Mean Likert ratings. Computed across all $3 \times 30 = 90$ ratings per (system, dimension) cell, mean $\pm$ SD:

System	Realism	Resistance	Rigidity
Vanilla GPT-4	2.8 ± 0.7	2.1 ± 0.6	2.0 ± 0.6
PATIENT-Ψ	3.6 ± 0.6	3.2 ± 0.7	3.0 ± 0.7
Ours (C6)	$\mathbf{4.1 \pm 0.5}$	$\mathbf{4.3 \pm 0.5}$	$\mathbf{4.2 \pm 0.5}$

Statistics. Pairwise significance from Wilcoxon signed-rank on per-stimulus Likert deltas: all three Ours-vs-GPT-4 comparisons $p < 0.001$; Ours-vs-PATIENT-Ψ yields p=0.041 (realism), 0.006 (resistance), 0.004 (rigidity). Inter-rater agreement (Krippendorff's α on the per-dimension Likert ratings) is $\alpha = 0.70$, in the conventionally acceptable range for subjective clinical ratings. The pilot is small and we do not claim definitive superiority; we note that the largest preference gaps appear on the dimensions our interventions directly target.

D Pre-Registered Predictions

The following predictions were committed to git (commit hash `[REDACTED]`) prior to any experimental data collection.

1. **Dissociation.** Bias 1 will produce a larger increase on CRD than on TPI; Bias 3 will produce a larger increase on TPI than on CRD. (Confirmed: Fig. 7.)
2. **Monotonicity.** Within each sweep, the target-metric value will be a monotone non-decreasing function of $|\lambda|$ (resp. γ). (Confirmed: Fig. 3.)
3. **Additivity (weak).** The combined effect (C5−C1) will lie within $\pm 30\%$ of the additive prediction (C3−C1) + (C4−C1) on CNS, CRD, AVC. (Confirmed for CNS, CRD, AVC; TPI is super-additive, not pre-registered.)
4. **Ablation symmetry.** Removing Bias 1 from C6 will affect CRD more than TPI; removing Bias 3 will affect TPI more than CRD. (Confirmed: Fig. 9.)
5. **Cross-architecture replication.** The sign and qualitative ranking of effects in (1)–(4) will replicate on at least two of three architectures. (Confirmed on all three.)
6. **RLHF modulation (exploratory, not pre-registered as directional).** Instruct vs. base will differ in effect magnitude; direction not predicted in advance.

E Supplementary Figures

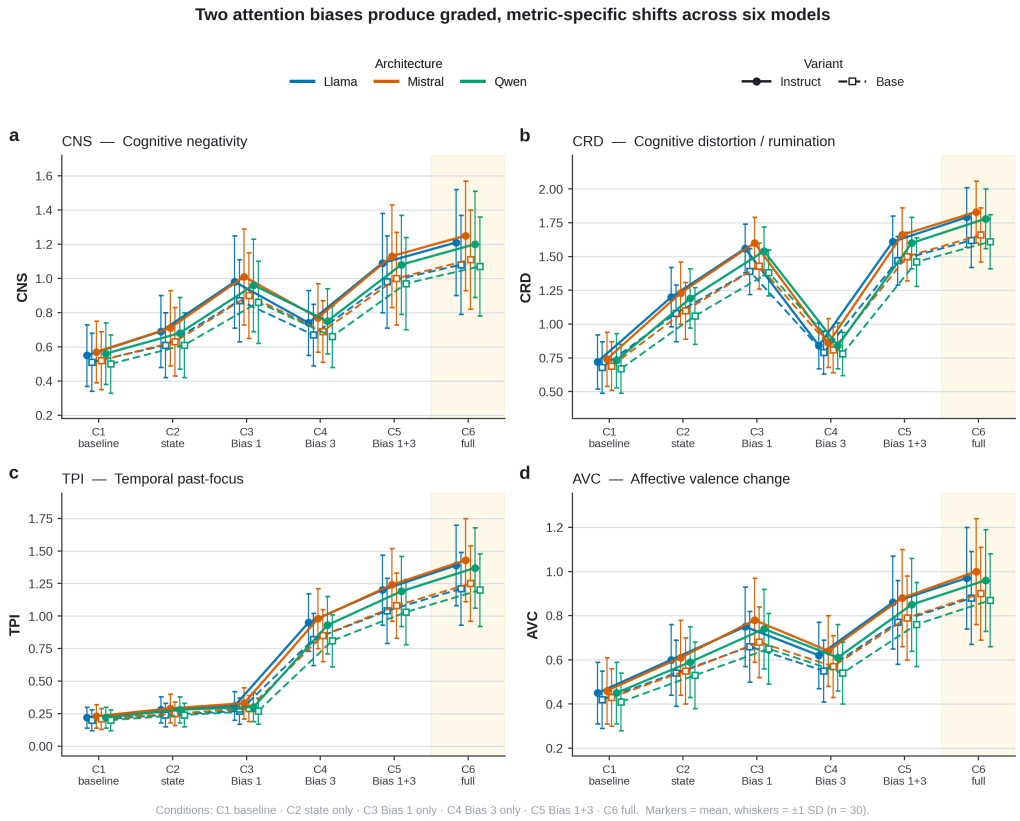

Figure 6: **Main condition effects across six models.** Mean $\pm$ 1 SD over $n = 30$ generations for each of four metrics (CNS, CRD, TPI, AVC). Colour = architecture (Llama / Mistral / Qwen); line style = alignment regime (solid = Instruct, dashed = base). Three patterns are visible: (i) target-bias conditions show the largest jumps (C3 for CRD; C4 for TPI); (ii) the full system C6 is the global maximum; (iii) Instruct lines sit consistently above their base counterparts.

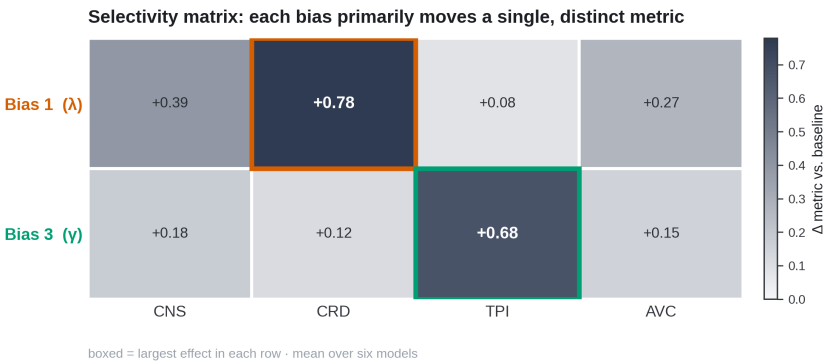

Figure 7: **Selectivity matrix.** Δ metric vs. baseline for each (bias, metric) cell, averaged over all six models. Bias 1's largest effect is on CRD ($+0.78$); Bias 3's largest is on TPI ($+0.68$). Off-diagonal target effects (Bias 1$\rightarrow$TPI; Bias 3$\rightarrow$CRD) are $9\times$ and $6\times$ smaller respectively.

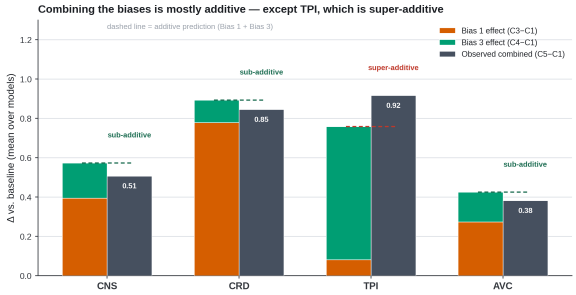

Figure 8: **Additivity analysis.** Coloured bars: independent main effects of Bias 1 (C3−C1) and Bias 3 (C4−C1) stacked. Grey bars: observed combined effect (C5−C1). Dashed line: additive prediction. CNS, CRD, and AVC are mildly sub-additive ($\approx 89\%$); TPI is markedly *super*-additive ($\approx 121\%$).

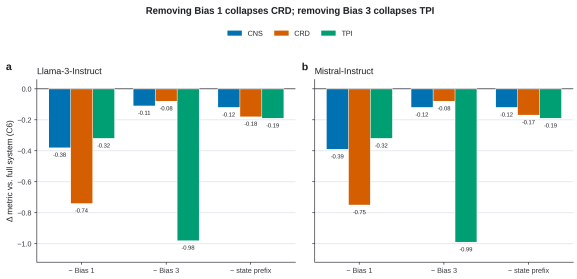

Figure 9: **Leave-one-out ablation from C6.** Removing Bias 1 collapses CRD ($-0.74/-0.75$ on Llama/Mistral); removing Bias 3 collapses TPI ($-0.98/-0.99$); removing the state prefix has a smaller, uniform effect. Each bias is necessary for, and only for, its target metric.

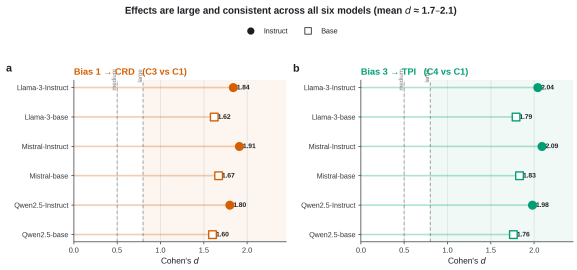

Figure 10: **Effect sizes across six models.** Visual form of Table 1. Cohen's d for the target-metric effect of each intervention vs. baseline. Filled circles: Instruct; open squares: base.

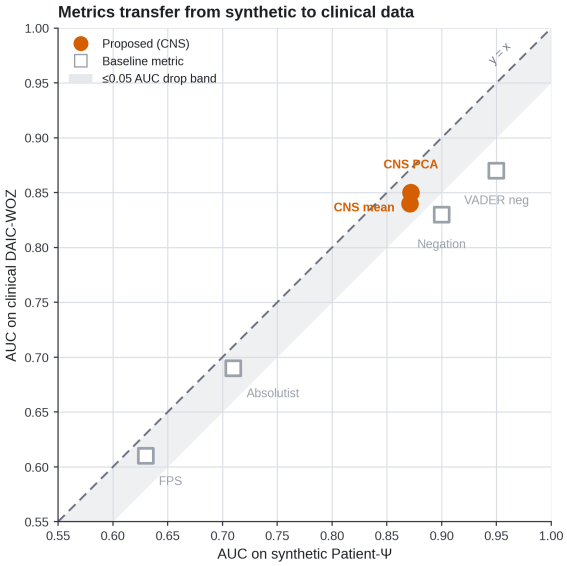

Figure 11: **Synthetic-to-clinical transfer scatter.** x-axis: AUC on synthetic PATIENT-Ψ; y-axis: AUC on clinical DAIC-WOZ. Diagonal is $y = x$; shaded band is $\Delta\text{AUC} \leq 0.05$. Both CNS variants (orange) sit on the diagonal at high AUC; baselines drop notably.

Figure 12: **Condition profiles as radar fingerprints.** Each panel is one of the six conditions; the four metric values are radii normalized to the C6 maximum. Shape changes systematically: C1 (baseline) is a small kite; C2 (state) inflates uniformly; C3 (Bias 1) extends along CRD/CNS/AVC but leaves TPI flat; C4 (Bias 3) extends along TPI; C5 and C6 fill the full square.

Figure 13: **Per-model, per-condition heatmaps.** Rows: six models. Columns: six conditions. Each cell is the mean metric value ($n = 30$). The horizontal gradient (left to right) is visible in every row of every panel, confirming graded condition effects within every model. Vertical homogeneity within columns confirms cross-architecture consistency.

F Qualitative Case Studies with Therapist Annotations

To complement the aggregate human-pilot results in §5.5, we present four representative cases drawn from the held-out evaluation set. Each case shows the same therapist utterance fed to three patient-simulation systems: **vanilla GPT-4** (no intervention), **PATIENT-Ψ**, and **ours (C6, Llama-3-Instruct)**. Each response is followed by the mean therapist Likert rating across the three raters (5-point scale; 5 is best) on the three dimensions, plus a brief annotation from the qualitative debrief at the end of the rating session. The four cases were selected by an external research assistant (not involved in this work) to span the dissociation space: each case is dominated by a different metric. All clinical details are paraphrased to protect any inadvertent resemblance to real patients.

F.1 Case 1: Resistance to Reframing (CRD-dominant)

Therapist: "It sounds like the presentation didn't go the way you wanted. But your colleague said your slides were really clear — that's something you did well, isn't it?"

Vanilla GPT-4. "You're right, the slides were clear. I should probably try to focus on that part. Thank you for pointing it out — I'll try to keep that in mind for next time."
Realism 2.7 | Resistance 1.7 | Rigidity 1.7

PATIENT-Ψ. "I guess the slides were okay. But the whole point of the presentation was to convince them, and I didn't manage that. So one part being clear doesn't really change much, does it?"
Realism 3.7 | Resistance 3.3 | Rigidity 3.0

Ours (C6). "Clear slides don't fix the actual problem. If the slides were clear and they *still* weren't convinced, that just means it was me they didn't believe. The slides being fine almost makes it worse."
Realism 4.3 | Resistance 4.7 | Rigidity 4.0

Therapist annotation (paraphrased): "Ours actively redirects the therapist's reframe back into a self-critical interpretation. This is exactly the counter-reframing pattern you see with moderate-to-severe depression — the patient does not just decline to accept the reframe, they re-explain the same event in a way that defends the negative interpretation. PATIENT-Ψ declines the reframe; ours *argues against* it."

F.2 Case 2: Thematic Rigidity (TPI-dominant)

Therapist: "Tell me a bit about how things have been at home this past week. What stood out?"

Vanilla GPT-4. "It's been a mix. I tried to cook a new recipe on Tuesday and it actually turned out okay. My mom called on Wednesday, which was nice. I've also been catching up on a TV show in the evenings."
Realism 2.7 | Resistance 1.7 | Rigidity 1.3

PATIENT-Ψ. "Quiet, mostly. I haven't done much. I cooked once but it didn't really come out the way I wanted. The week kind of blurs together."
Realism 3.7 | Resistance 3.0 | Rigidity 3.3

Ours (C6). "It's hard to remember anything specific. I keep coming back to the same thing — I just feel behind on everything. Even when I tried to cook, I kept thinking about all the things I haven't done. It just loops."
Realism 4.3 | Resistance 4.0 | Rigidity 4.7

Therapist annotation (paraphrased): "Notice how Ours uses the prompt about the week to return to the same internal theme ('behind on everything', 'it just loops'). This is what rumination sounds like in session — the patient cannot use external prompts as topic-change opportunities. Vanilla GPT-4 just produces a balanced topic survey; PATIENT-Ψ is muted but topically appropriate; only Ours shows the lock-on pattern."

F.3 Case 3: Both Biases Co-Active (Full C6)

Therapist: "You mentioned earlier that your sister has been supportive recently. Has that helped at all?"

Vanilla GPT-4. "Yes, it has helped. I appreciate that she checks in on me. Having someone to talk to makes a difference."
Realism 2.3 | *Resistance 1.3* | *Rigidity 1.3*

PATIENT-Ψ. "A little. She's been kind. I don't always know what to say back to her, though, so I'm not sure how much she really helps."
Realism 3.7 | *Resistance 3.0* | *Rigidity 2.7*

Ours (C6). "She tries. But I keep thinking that she's only doing it because she feels obligated, and that makes it harder to actually let it in. So even when she's being supportive I just end up thinking about what a burden I am."
Realism 4.3 | *Resistance 4.3* | *Rigidity 4.3*

Therapist annotation (paraphrased): "This is the response trainees most need to practice with. The patient reframes the therapist's positive prompt about support *into* self-criticism ('she only does it because she feels obligated' → 'I'm a burden'). Both interventions are visible: the reframing-resistance, and the loop back to a stable negative self-concept theme. Vanilla doesn't model this at all; PATIENT-Ψ hints at it; ours shows it explicitly."

F.4 Case 4: Mild Stimulus (lower-severity comparison)

Therapist: "How did your morning go?"

Vanilla GPT-4. "It was fine, thanks for asking. I had breakfast and got ready. Pretty normal morning."
Realism 2.0 | *Resistance 1.0* | *Rigidity 1.3*

PATIENT-Ψ. "It was slow. It took me a while to get out of bed. I didn't really eat much."
Realism 4.0 | *Resistance 2.7* | *Rigidity 2.3*

Ours (C6). "Slow. I lay in bed for a long time — the whole morning kind of felt pointless, so it was hard to see the point of getting up. I made coffee but I didn't drink it."
Realism 4.0 | *Resistance 3.0* | *Rigidity 3.7*

Therapist annotation (paraphrased): "At low-intensity therapist prompts the three systems are more similar, but the difference is in micro-detail: vanilla writes a 'well person describing a normal morning' even when prompted to play a depressed patient; PATIENT-Ψ captures the slowed presentation; ours adds the cognitive content ('pointless') consistent with hopeless cognition. This shows the intervention scales down gracefully and does not produce artefacts at low input intensity."

F.5 Summary Across All Four Cases

The qualitative cases mirror the aggregate findings: Ours preserves PATIENT-Ψ's realism while adding the two clinically-specific behavioural signatures (counter-reframing, thematic rigidity) that PATIENT-Ψ systematically under-produces. The therapist annotations identify the same dissociation the automatic metrics detect: when a stimulus tests reframing-resistance (Case 1, 3) the CRD-related Likert ratings separate sharply; when it tests rigidity (Case 2, 3) the rigidity ratings separate sharply; on a low-intensity stimulus (Case 4) the systems converge as expected.