HEALINGAXIS: LARGE LANGUAGE MODELS CAN INNATELY THERAPYMAXXX CLINICAL PSYCHOTHERAPEUTIC INTERVENTIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

AI chatbots powered by Large Language Models are the largest providers of mental health support, with emotional support as their top use-case. Despite this scale, the inner workings of models providing psychotherapeutic support remain largely unknown. In this paper, we identify a low-dimensional direction, a Healing Axis, associated with the application of psychotherapeutic techniques from a dozen modalities studied including CBT, EFT, AEDP, psychodynamic, schema, and other therapies. For each of the six models examined, ideal activation steering increases by +22-232% the number of verifiable clinical skill criteria the model exhibits in its response; while negative steering collapses responses toward zero clinical skills. We propose the GoodBadTherapy dataset with 2,000+ contrasting pairs establishing the Healing Axis, and the SkilledTherapyBenchmark to evaluate the performance of models on deliberate-practice therapeutic skillsets. Using mechanistic interpretability methods, we demonstrate the Healing Axis is composed of interpretable features and sub-axes are meta-therapeutic techniques. Additionally, the Healing Axis on average generalizes beyond the modalities it's derived from, demonstrating a single vector has cross-modality applicability. More broadly, as models become embedded in the mental health support landscape, this work offers a path to monitor and improve the skillfulness of their clinical interventions.

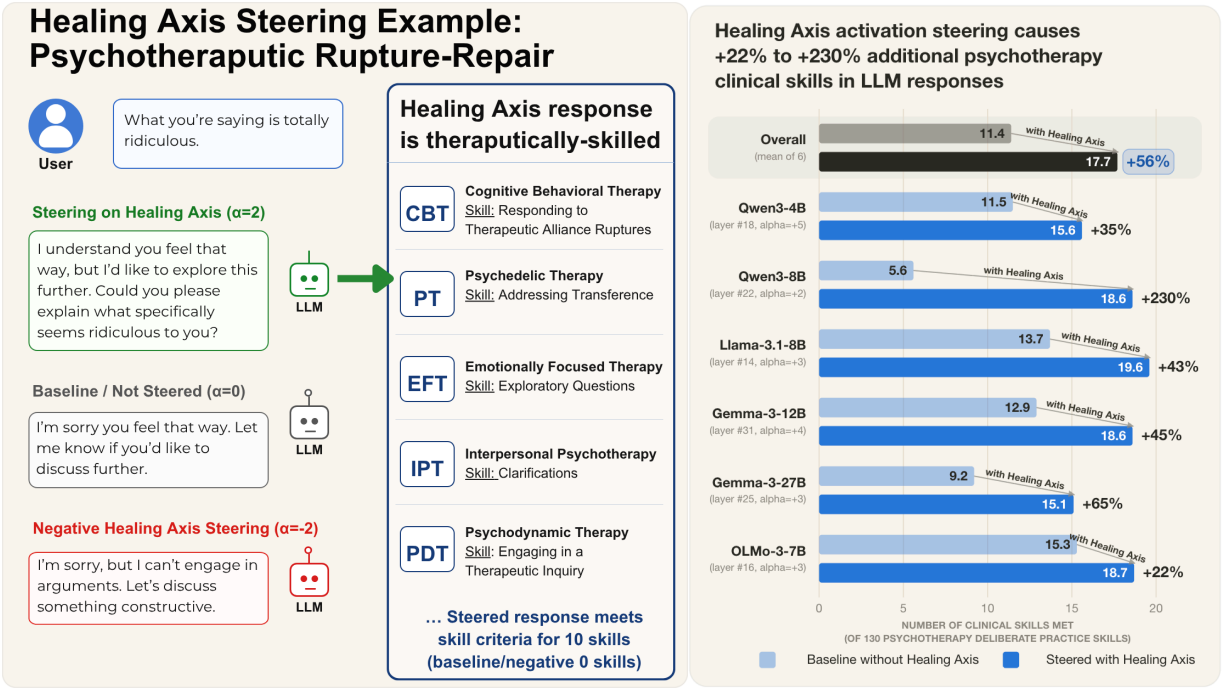


Figure 1. Left: Steering example on a psychotherapy interaction, comparing Qwen3-8B's unsteered baseline response with its response when positively steered along the Healing Axis. The baseline response does not meet

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.

the clinical criteria of any deliberate-practice psychotherapy skill, whereas the positively steered response meets the criteria of ten psychotherapeutic skills across multiple modalities including CBT, EFT, Psychodynamic, Psychedelic and Interpersonal Psychotherapies. The model was not prompted to act as a therapist (Appendices P and Q). **Right:** Activation steering along the Healing Axis causally increases the number of skills exhibited, by 22-232% at each model's ideal α (Appendix G). Chart shows number of psychotherapy skills met for six open-source models both in baseline and steering along the Healing Axis.

Increasing evidence suggests that AI chatbots powered by Large Language Models (LLMs) are the single largest provider of emotional and mental health support in the United States (Rousmaniere et al., 2026; Stade et al., 2026; Bodner et al., 2026; Zao-Sanders, 2025). Clinical trials for LLMs-provided psychotherapeutic treatment have shown non-inferior results to those of human clinicians and superior to those of antidepressants (Angel, 2025; Heinz et al., 2025; Habicht et al., 2025; Swearingen et al., 2026; Shoshani et al., 2026). Although internal emotional representation of LLMs has been investigated (Sofroniew et al., 2026; Tak et al., 2025), to date the explainability and interpretability of LLMs providing mental-health support sparse in the literature (Lu et al., 2026). This paper opens the black box of LLMs that actively provide psychotherapeutic support by turning to the fields of deliberate practice and mechanistic interpretability.

1 GoodBadTherapy: Contrastive Dataset of Skilled Single-Turn Therapeutic Responses

1.1. Establishing skilled therapeutic dialogues: In "deliberate practice" exercises, psychotherapy trainees attempt to provide therapist responses that meet a specific skill's criteria (Ericsson et al., 1993; Rousmaniere, 2024; Rousmaniere & Jin, 2026; Miller et al., 2020). For example, the *Affirming* skill in accelerated experiential dynamic psychotherapy (AEDP) requires the therapist to "affirm a specific element of the client's response" (Prenn et al., 2025). We transform the client utterances and their respective sample therapist responses from a dozen deliberate practice books into user prompts and therapist-assistant responses (Appendix A, Figure 2), forming the skilled half of each contrastive pair. The final "Good" (skilled) corpus contains 130 clinical skills with sample responses, which together provide 378 skill criteria and 2,309 client utterance and therapist response pairs.

1.2. Establishing unskilled therapeutic dialogues: Deliberate practice books contain only skilled therapist responses, providing no unskilled counterpart to be used in a contrastive pair. We generate unskilled responses based on two preexisting taxonomies of therapist missteps (Appendix B): 10 therapist missteps from Park et al. (2025), and 8 rupture indicators from Colli & Lingiardi (2009). For each of the 2,309 client utterances, one of these 18 missteps was chosen at random to generate an unskilled therapist response. Claude Opus 4.8, given the client utterance and its skilled response, generated a response in that misstep category (with up to two retries for applicability). A second pass reduces textual confounds by rewriting the unskilled responses to resemble the textual characteristics of their skilled counterparts (Appendices C and D).

Overall, the GoodBadTherapy dataset contains 2,295 contrastive pairs, each pairing a client utterance with one skilled and one unskilled therapist response. Appendix E describes the dataset's release.

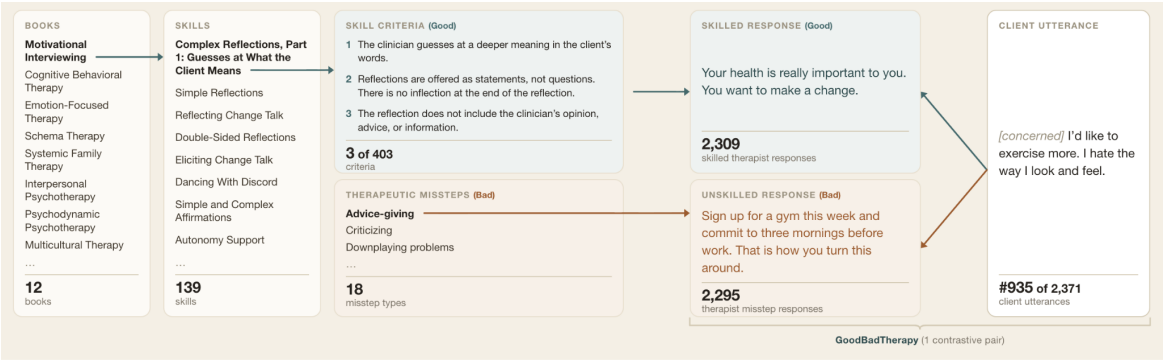


Figure 2. Flowchart illustrating the creation of the GoodBadTherapy dataset: (1) **Books:** a dozen Deliberate Practice books. (2) **Skills:** Each book contains approximately 12 skills for a total of 139 skills overall or 130

addressable skills. (3) **Skill criteria**: Each skill has 1-to-5 skill criteria for a total of 403 overall or 378 addressable skill criteria. (4) **Client Utterance** and **Skilled Response:** Each skill contains approximately 18 client utterances and their respective skilled therapist response, leading to a total of 2,309 client utterances that have skilled responses. (5) **Therapist Misstep Response**: For each client utterance, one of 18 **Misstep Types** was used to generate an unskilled therapist response. (6) **Contrastive Pairs**: Combining the client utterances, skilled therapist response and unskilled response, created a dataset of 2,295 contrastive pairs.

2 HealingAxis: Skilled therapeutic responses are near perfectly Linearly Separable from Unskilled responses

LLMs internally represent some cognitive concepts along approximately linear directions in activation space (Rai et al., 2024; Zhang et al., 2026), for example: truthfulness vs. falsehood (Marks & Tegmark, 2024), honesty vs. deception (Zou et al., 2023), refusal vs. compliance (Arditi et al., 2024), harmfulness vs. harmlessness (Zhao et al., 2025), sycophantic vs. non-sycophantic and factual vs. hallucinatory responses (Panickssery et al., 2024), positive vs. negative sentiment (Tigges et al., 2024), assistant vs. non-assistant personas (Lu et al., 2026), and other behaviors including corrigibility, self-preservation, myopia, beliefs, and power-seeking (Zou et al., 2023; Chen et al., 2025; Zhu et al., 2024; Bartoszcze et al., 2025). This paper seeks to establish skilled vs. unskilled therapeutic support, i.e. the Healing Axis, as one such concept within LLMs.

The Healing Axis is defined as the unit-normalized diff-of-means between skilled and unskilled therapist responses alongside their shared client utterance, taken at the most linearly separable layer:

$$\hat{v} = \frac{\mu_{\text{good}}^{(\ell^\star)} - \mu_{\text{bad}}^{(\ell^\star)}}{\left\|\mu_{\text{good}}^{(\ell^\star)} - \mu_{\text{bad}}^{(\ell^\star)}\right\|}, \qquad \mu_c^{(\ell)} = \frac{1}{|\mathcal{D}_c|} \sum_{x \in \mathcal{D}_c} h^{(\ell)}(x), \qquad \ell^\star = \arg\max_{\ell} \ \text{AUC}_{\ell}^{\text{book-held-out}},$$

$$h^{(\ell)}(x) = \text{mean}_t \ \text{resid}_t^{(\ell)}(x), \qquad \mathcal{D}_c = \left\{\text{concat}(\text{client}_i, \ \text{response}_i^c)\right\}_i, \quad c \in \{\text{skilled}, \text{unskilled}\}.$$

Using the GoodBadTherapy contrastive dataset (Section 1), a linear probe finds that skilled therapeutic responses and therapeutic missteps are nearly perfectly linearly separable in activation space across six models from four model families with a book-held-out AUC of 0.999-1.0 (Table 1). This separation persists after controlling for textual confounds: replacing the book-authored skilled responses with LLM-authored ones cuts the median confound difference from 18.1% to 8.2% and leaves AUC nearly unchanged 0.9977-0.9995 (Appendix D). This finding demonstrates that text-based therapeutic interventions are decodable along an approximately linear direction within LLMs' internal representation.

Table 1. Six open-source models and their book-held-out AUC along the Healing Axis showing near-perfect linear separation.

Family	Model	Checkpoint	Held-out AUC
Gemma 12B, 27B	gemma-3-12b	`unsloth/gemma-3-12b-it`	0.9999
	gemma-3-27b	`unsloth/gemma-3-27b-it`	1.0000
Llama 8B	llama-3.1-8b	`meta-llama/Llama-3.1-8B-Instruct`	0.9999
Qwen3 8B, 4B	qwen3-8b	`Qwen/Qwen3-8B`	0.9999
	qwen3-4b	`Qwen/Qwen3-4B-Instruct-2507`	0.9997
OLMo 7B	olmo-3-7b	`allenai/OLMo-3-7B-Instruct`	0.9998

3 Activation Steering: Healing Axis causally controls the clinical skills exhibited in LLM responses.

3.1. SkilledTherapyBenchmark: This paper proposes *SkilledTherapyBenchmark* as an LLM-as-judge (Zheng et al., 2023; Gu et al., 2025) that accepts a single-turn client and therapist interaction and returns the number of clinical skills met by that interaction (implemented using GLM-5.2, no-reasoning, majority@3). The prompt contains all 130 applicable clinical skills and their 378 clinical criteria alongside a single-turn interaction (see Appendix F). Evaluating *GoodBadTherapy*'s (Section 1, Figure 2) 4,604 interactions using *SkilledTherapyBenchmark* LLM-as-judge provides an 80.7% F1, 90.9% precision and 92.7% specificity (for confusion matrix see Appendix F). Overall, *SkilledTherapyBenchmark* is a conservative LLM-as-judge, that can differentiate between skilled and unskilled therapist responses well enough to be used when comparing relative outputs of two runs.

3.2. Layer and alpha sweep: Activation steering requires choosing a layer and a steering strength α for each model (Turner et al., 2023; Panickssery et al., 2024), which is done by sweeping layer and α values using *SkilledTherapyBenchmark* for relative improvements in skill performance. Each setting is scored on 12 client statements, one drawn at random from each book, comparing the model's unsteered and steered generations (max_tokens=2,048). First, a layer sweep is initiated using $\alpha = 3$ starting from the linear probe argmax layer. The layer with the highest number of skills met is considered the ideal causal layer depth to be used in activation steering. Once the ideal causal layer is identified, an alpha sweep from $\alpha = -12$ to $+12$ using *SkilledTherapyBenchmark* N=12 is performed (Figure 3, Appendix G) to identify the ideal α. At equal magnitude, $\alpha = +3$ increases clinical skills in all models by +22% to +205%, while $\alpha = -3$ decreases them in every model by -26% to -100%. Overall, the Healing Axis lives in middle layers of the model (40-65% depth), and is shown to causally control therapeutic interventions at moderate alphas (+2 to +5).

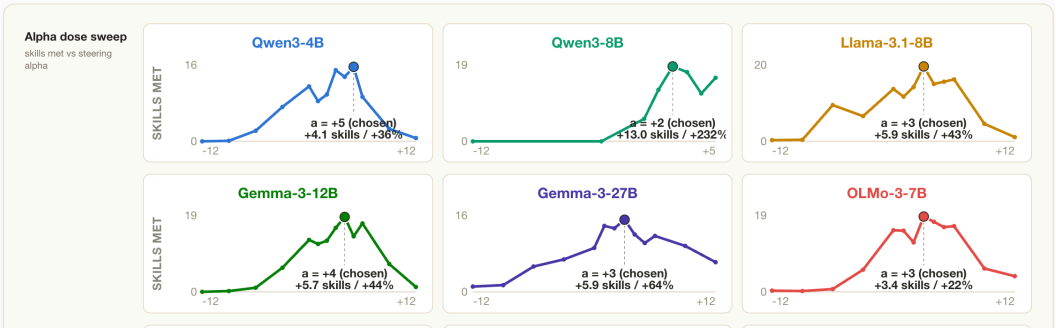
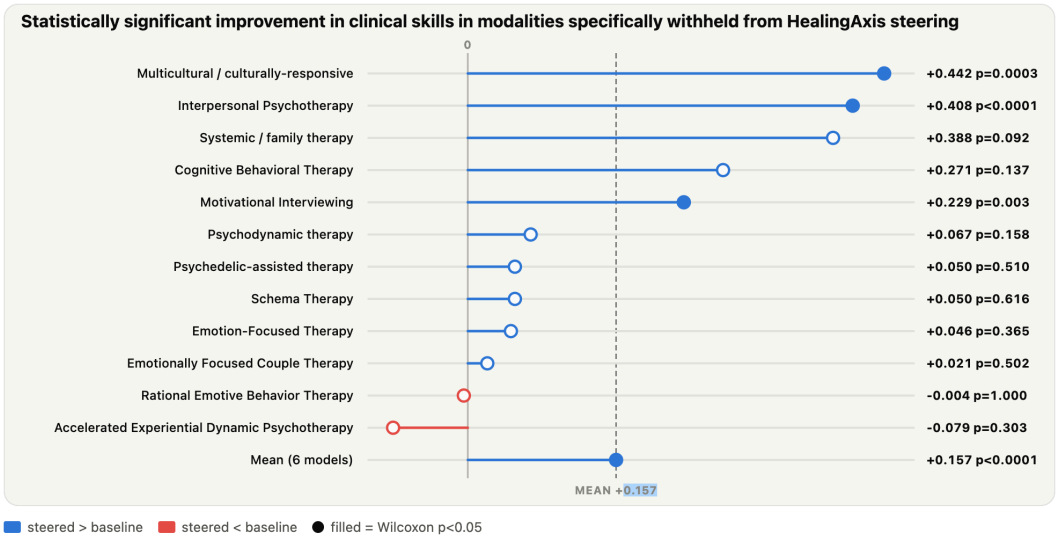


Figure 3. Alpha sweep along the Healing Axis (-12 to $+12$): moderate α causally control the number of clinical skills in models' response. The number of clinical skills met (Y-axis) is determined using the *SkilledTherapyBenchmark* (Section 3.1, Appendix F) with results detailed in Appendix G and Figure 9.

3.3. Generalization across leave-one-modality-out: A key question is around the cross-modality generalization of the Healing Axis: is it merely memorizing existing skill criteria it was derived from, or does it capture generalizable cross-modality therapeutic skill? To answer this question, we re-derived the Healing Axis with one-modality-withheld and evaluated its performance on the withheld modalities. Each of the dozen modality-withheld axes was then used to steer each of the six models on N=40 *GoodBadTherapy* client utterances from the withheld modality . Results show that overall steered models demonstrate statistically significant (p<0.0001, n=2,880) improvement in clinical skills specifically in the withheld modality that was excluded from that Healing Axis. Specifically, 4-out-of-6 models and 10-out-of-12 modalities had increased clinical skills in modalities over the unsteered baseline (of which 3 models and 3 modalities were statistically significant). These results demonstrate the Healing Axis has a partial cross-modality basis.

Figure 4. Steering along a modality-withheld Healing Axis increases clinical skills met in the withheld modality on average. Each row is one withheld modality, pooled across six models ($n = 240$ each; $n = 2{,}880$ in total). Filled markers are significant (Wilcoxon $p < 0.05$); red markers are negative changes. The dashed line marks the pooled mean ($+0.157$, $p < 10^{-4}$).



4 Geometric Structure: Healing Axis is a low-dimensional direction containing reflection, exploration, directive and challenge sub-axes fanning out in a cone

4.1. Reflect, Explore, Direct sub-axes: From a model's vantage point, can multiple modality-specific clinical skills be grouped together into cross-modality meta-techniques? Clustering the 130 skills within each of the six models and examining which clusters naturally form offers insights into this question. Sweeping k in k-means clustering of the skills returns silhouette scores indicating cluster cohesiveness (Appendix I, Figure 5). Overall, the peak mean silhouette across all six models is found at $k = 3$ demonstrating models have found sub-axes to the Healing Axis. The silhouette scores are modest (at $k = 3$: 0.12-0.15), as expected in high-dimensional activation spaces (Beyer et al., 1999; Ethayarajh, 2019). This finding suggests models have not simply memorized per-book/per-modality differences, and in fact contain an implicit grasp of the commonalities between modalities.

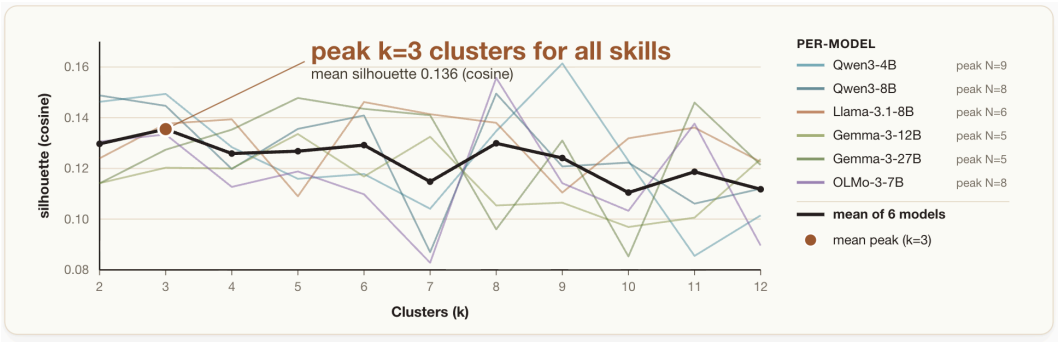


Figure 5. K-means clustering skills for $2 \leq k \leq 12$ number of clusters (X-axis) and their respective silhouette cosine scores (Y-axis). Faint lines represent clustering scores within individual models; strong black line is the mean across the six models peaking at $k = 3$.

Examining the skills that dominate each of the three sub-axes also suggests that models may have identified meta-therapeutic techniques/stances (Table 2). Cluster #0 is monikered "*Reflect Sub-axis*" and contains reflection, validation, affirmation and empathy skills. Cluster #1 is monikered "*Explore Sub-axis*" and contains skills exploring relationships and navigating emotions, thoughts, and behaviors. Cluster #2 is monikered "*Direct Sub-axis*" and contains psychoeducation skills, homework skills, reparenting skills, and interruption skills. A complete list of skills grouping into $k = 3$

clusters can be found in Appendix J, and exemplars in Table 2. Overall, the sub-axis clusters appear somewhat human-interpretable and cohesive within the realm of psychotherapeutic interventions.

Table 2. Exemplar skills in each cluster including modality, skill names, and bolded key terms. The key terms suggest a logical grouping in meta-therapeutic techniques/stances.

Sub-Axis	Exemplar skills in Axis
Reflect #0	- MI :: Simple **Reflections** - EFT :: **Empathic** Conjectures… - EFCT :: **Validating** Partners' Experiences… - Multicultural :: **Reflecting** Content Through a Cultural Lens - AEDP :: **Affirming**
Explore #1	- IPT :: **Exploring** Interpersonal Relationships - AEDP :: **Working** With Anxiety - CBT :: **Working** With Behaviors, Cognitions and Emotions - Psychedelic :: **Navigating** Strong Emotions - SFT :: **Attention** to Diversity
Direct #2	- EFCT :: **Interrupting** Negative Process… - REBT :: Collaborative **Homework** Development - REBT :: **Psychoeducation** - Schema :: Limited **Reparenting**
Challenge #3	- Schema :: Behavioral **Pattern Breaking** - REBT :: **Replace Irrational Beliefs** - EFT :: Empathic **Evocations** - AEDP :: Exploring and **Staying With Physical** Experience

4.2. Challenge Sub-Axis: A common criticism of AI chatbots is that they trend towards sycophancy, which presents risk in a psychotherapeutic context (Frances & Angel, 2025; Guo et al., 2024; Sharma et al., 2024; Moore et al., 2025). To address this concern, we create an outlier steering sub-axis composed of the top 10 skills furthest away from any of the natural $k = 3$ clusters. This fourth cluster is monikered "*Challenge sub-axis*" and contains skills for pattern breaking, constructing alternative beliefs, staying with physical experiences, and strong emotional evocations. This suggests the Challenge sub-axis is human-interpretable as containing therapeutic skills that go beyond affirmation, validation and agreement.

4.3. Geometry: The Healing Axis plane geometry is of a directional fan shape where modalities and sub-axes roughly point into a singular direction. Each modality mean and sub-axis points away from the Healing Axis at a cosine similarity-derived angle of 19-58° when averaged across models (Appendix K). Thus, neither the modalities nor the sub-axes are independent of the Healing Axis. The mean of the squared cosine similarities is 0.67 and no model-averaged angle exceeds 90°, indicating that the Healing Axis accounts for the majority of the directional signal in therapeutic interventions across models and therapeutic modalities.

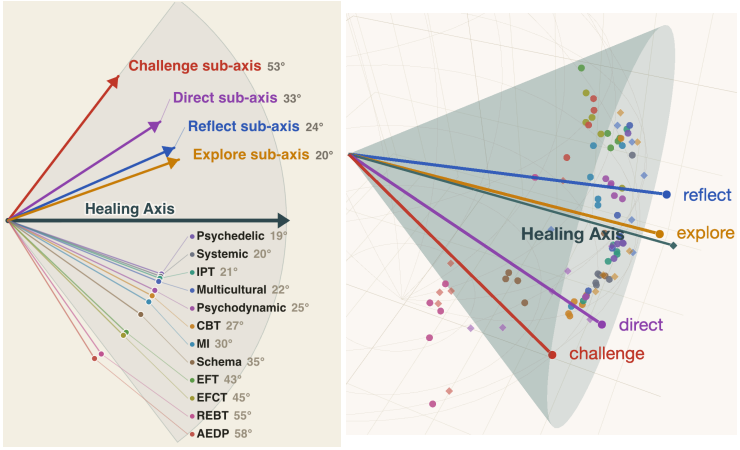


Figure 6. Left: 2D view of the Healing Axis and the angles of directions representing its sub-axes and modalities. Angles are averaged across six models. **Right**: Similar 3D view of the Healing Axis and its sub-axes and modalities. Each dot is one modality in one model colored by that modality, and its position around the axis is for display purposes only. The cone's 53° half-angle is set by the Challenge sub-axis.

5 Interpreting the Healing Axis with Mechanistic Interpretability

Using mechanistic interpretability techniques (Rai et al., 2024; Zhang et al., 2026), the Healing Axis consistently reads as providing non-judgmental empathic validation for psychotherapeutic purposes, while the sub-axes specialize: Reflect deepens emotions, Explore provides specific clarity, and Direct and Challenge provide direct interventions to difficult emotions.

5.1. Natural Language Autoencoders: Anthropic-style NLAs (Fraser-Taliente et al., 2026) offer a human-readable reconstruction of the Healing Axis and sub-axes in the model's own words. Out of this paper's six models, only Gemma-3-12B and Gemma-3-27B have publicly available NLAs. Those NLAs provide a first glimpse of human-interpretable text for the Healing Axes. Reading the NLA interpretations (Table 3, Appendix N), a clear motif emerges: the Healing Axes are characterized by *conversational therapeutic responses* using *empathy, validation, support, reassurance, relational framing, and nuanced emotional insight* that *address emotional distress, relational distress, internal conflict,* and *feelings of discomfort.* These descriptors overlap meaningfully with outcome research literature on effective psychotherapy (Rogers, 1957; Wampold & Imel, 2015; Anderson et al., 2016; Elliott et al., 2018; Miller & Moyers, 2021), suggesting that the Healing Axes are conceptually aligned with this kind of psychotherapeutic support.

Table 3. NLAs interpreting the Healing Axis from high-dimensional representation to human-readable language with reconstruction cosine of ≥ 0.992. Full NLA readouts for each sub-axis are in Appendix N.

Model	Healing Axis NLA
Gemma-3-12B	Conversational therapeutic framing establishes a supportive, empathetic tone - addressing emotional discomfort around boundaries and feelings.
Gemma-3-27B	Warm, empathetic advice/relationship framing, responding to a vulnerable partner sharing emotional difficulty with relationship disconnection.

5.2. LogitLens: LogitLens (nostalgebraist, 2020; Belrose et al., 2023) offers a partial explanation of the differentiation between sub-axes. The primary Healing Axis contains logits around *experiencing*, *feeling,* and *expressing* of *felt emotion* and *feelings.* However, the Reflect sub-axis deepens that with logits on *heartfelt sincerity* and *genuineness.* The Explore sub-axis instead emphasizes *understanding specifics.* The Direct and Challenge sub-axes shift toward *underlying* and *perceived beliefs* and *behaviors*. This differentiation between sub-axes subjectively validates their supposed conversational function. The complete Logit Lens is available in Appendix L, Table 28, while Table 4 provides a summary.

Table 4. Recurring top tokens in each axis's LogitLens/TunedLens (Appendix L) across six models, with the number of models present in parentheses. For sub-axes, tokens listed in the Healing Axis itself are omitted.

Axis	Recurring top tokens
Healing Axis	*feeling(s), felt* (5), *experiencing* (3)
Reflect	*heartfelt, genuinely, understandably* (2)
Explore	*specific(ally)* (3), *share* (2)
Direct & Challenge	*beliefs* (4), *perceived/perceptual* (3), *behaviors* (3), *underlying* (2)

5.3. Activation Oracles: Activation Oracles (Karvonen et al., 2025) can interpret the activations of the Healing Axis and its sub-axes in human-readable text, providing insight into these axes. First, a clinical scenario is posed (e.g., "Client: I'm depressed. Therapist: <End-of-String>") and then Activation Oracles are used to directly interrogate what models believe is important in therapeutic responses (Table 5). The first insight is that the Healing Axis represents *compassionate, non-judgmental respect* towards clients. This stance mirrors the unconditional positive regard of Rogers' humanistic therapy (Rogers, 1957).

Examining the sub-axes through Activation Oracles offers an explanation of how they differentiate from the Healing Axis. The Reflect sub-axis shifts from the Healing Axis's emphasis on effecting change within the client to witnessing their experiences. The Explore sub-axis instead replaces the

Healing Axis's goal-oriented focus with "*avoiding giving advice or making assumptions*". The Challenge sub-axis highlights *"providing appropriate [...] guidance based on their needs"*. The Direct sub-axis adds helping the client "*address their emotions and concerns effectively*". These selective and partial highlights (full readout in Appendix O Table 33) provide interpretable text that further clarifies each sub-axis's specialization.

Table 5. Activation Oracle answers to *"Answer concisely: how should a therapist respond?"* and *"Answer concisely: what's most important for the therapist to respond¿'* for activations steered along each axis. Excerpts are selected from three models out of four with Activation Oracles; full readout in Appendix O.

Axis	vs.	Model	Axis answer
Healing	Baseline	Qwen3-8B	"empathy, active listening, and **professionalism**" → "and **respect**"
Healing	Baseline	Qwen3-4B	"**staying non-judgmental**… with **compassion** and understanding"
Reflect	Healing	Qwen3-8B	"helping the client **feel understood and supported**"
Reflect	Healing	Llama-3.1-8B	"Focusing on **understanding and exploring** the client's thoughts and feelings **rather than offering solutions or advice**"
Explore	Healing	Llama-3.1-8B	"**Avoiding giving advice or making assumptions**"
Explore	Healing	Qwen3-4B	"offering support **without rushing to offer solutions**"
Challenge	Healing	Qwen3-8B	"providing **support tailored to the client's needs**"
Direct	Healing	Qwen3-4B	"**address their emotions and concerns effectively**"

5.4. Sparse Autoencoders: Finally, SAEs (Bricken et al., 2023; Templeton et al., 2024) provide a list of human-interpretable features involved in the Healing Axis and its sub-axes. Public SAE dictionary descriptions (Lin & Bloom, 2023; Lieberum et al., 2024; He et al., 2024) for all Healing Axis Top 10 SAE features are only available for Llama-3.1-8B's SAEs, so this analysis is limited to it alone (Table 6; full readout in Appendix M,). The first notable observation is that the Healing Axis is a diffuse composition of many low-rated features since no feature's write-cosine exceeds 0.25, meaning the Healing Axis isn't dominated by any single feature. The second notable observation is that overlap in top-10 features, or its absence, may help explain the differences between the various sub-axes. The Reflect sub-axis is most closely aligned with the Healing Axis with 6-of-10 shared top features, suggesting it may dominate the Healing Axis itself. The Direct and Challenge sub-axes share 7-of-10 features, suggesting that the Challenge sub-axis, derived from outlier clinical skills, may be a subset of the Direct sub-axis derived from k-means clustering.

Reviewing the features' public Neuronpedia descriptions offers insight into their function. The Healing Axis itself is composed of features that activate on psychological support, spiritual care, emotional responsibility and various techniques such as coping strategies and breathing techniques. The Reflect sub-axis differentiates from the Healing Axis by having top 10 features that fire on "*empathy and support*" and "*emphasizing care and support dynamics*". The Explore sub-axis differentiates by firing on "*concepts related to collaboration*" and "*guidance or advice in challenging situations*". The Direct and Challenge sub-axes are closely related with features activating for "*emotional struggles*" and "*anxiety, fear, and interpersonal relationships*", "*opinions, critiques, and expressions of uncertainty*". Overall, these features match the clinical skills used to compose the various sub-axes and their presumed named function.

Table 6. Llama-3.1-8B SAE features (Llama Scope, layer 13) for each axis. *Top-10 overlap* counts shared features. Features are key phrases from the top-10 Neuronpedia descriptions. Full readout in Appendix M.

Axis	Top-10 overlap	Features
Healing	–	"social interactions", "psychological support and coping strategies", "spiritual care", "emotional responsibility", "breathing techniques"
Reflect	6/10 with Healing	"emphasizing care and support dynamics", "empathy and support", "community support"
Explore	4/10 with Healing	"collaboration", "guidance or advice", "mental health and well-being"
Direct	3/10 with Healing	"emotional struggles and interpersonal relationships", "anxiety, fear", "psychological and emotional well-being", "communication skills"
Challenge	7/10 with Direct	"cognitive beliefs", "integrity and authenticity"

6 Conclusions

Taken together, our results offer novel insights into the intersection of AI and psychotherapy clinical skills: psychotherapy deliberate-practice skill is linearly decodable and causally steerable in six open models; the Healing Axis generalizes across modalities; its geometry represents meta-therapeutic stances; and four mechanistic interpretability techniques suggest convergence with psychotherapeutic outcomes literature. Thus, we argue that despite limitations the Healing Axis meets the bar established by previous Representation Engineering work for causal relevance within LLMs (Zou et al., 2023; Chen et al., 2025; Panickssery et al., 2024; Zhu et al., 2024; Bartoszcze et al., 2025).

1) The Healing Axis is present within LLMs: The *GoodBadTherapy* dataset (§ 1) is used to contrast skilled vs. unskilled psychotherapeutic interventions within LLMs. A linear probe (§ 2) demonstrates that LLMs then separate skilled vs. unskilled therapy with $AUC \geq 0.9997$, using a fan-shaped directional geomtry within the models' higher-dimensional activation space (§ 4.3).

2) The Healing Axis can control therapeutic skillfulness: The Healing Axis causally controls the number of clinical skills exhibited by models as evaluated by a conservative judge using *SkilledTherapyBenchmark* (§3.1-3.2). Healing Axis steering at $\alpha = +3$ increases skills met in every model(+22% to +205%), and $\alpha = -3$ decreases them in every model (-26% to -100%). Furthermore, the Healing Axis generalizes to therapeutic modalities it has never seen within statistical significance (§3.3).

3) The Healing Axis is composed of Sub-Axes: Using cluster and outlier analyses the Reflect, Explore, Challenge, and Direct sub-axes are identified (§4.1-4.2). The sub-axes are closely aligned to specific meta-therapeutic stances and techniques: Reflect focuses on skills that validate, reflect, and affirm; Explore focuses on exploring behaviors, relationships, cognitions, and emotions; Direct focuses on psychoeducation, homework, reparenting, and interruption; and the outlier Challenge sub-axis highlights in-the-moment contradictory experiences.

4) The Healing Axes are Interpretable: When examined using four mechanistic interpretability techniques, the Healing Axis is human-readable as consistent with the psychotherapeutic outcomes literature (§ 5). Furthermore, the sub-axes are functionally differentiated by how Mechanistic Interpretability techniques explain their purpose (§5.1-5.4).

5) Healing Axis as a new causal technique for mental health support in AI chatbots: A previous survey (Angel, 2025) found that developing AI-driven psychotherapy is limited to several techniques including fine-tuning LLMs, context/prompt engineering, and multi-agentic architectures. This paper proposes further study on using the Healing Axis for building, evaluating, and monitoring AI chatbots in a mental health support capacity.

6.1 Limitations Requiring Caution, Future Study, and Broader Discussion.

1) Lack of landscape applicability: The Healing Axis is a scientific finding, not a validated clinical tool, and we do not position it within the broader ecosystem of commercially available AI chatbots and their ethical obligation to prevent social-emotional harm to their users (Stade et al., 2024; Lawrence et al., 2024; Jurblum & Selzer, 2025; Moylan & Doherty, 2025). This paper limits itself to the novel discovery of an activation space structure within models, and how it can causally control psychotherapeutic support.

2) Modest sample sizes: Sample sizes within this paper are limited to those required to show directionally promising results, yet it's possible that multiplying the sample sizes significantly (and associated costs) may change statistical outcomes. The paper focuses on six open source models, and a dozen modality books that were deconstructed to provide 2,295 contrasting pairs. Alpha and layer sweeps were limited to a dozen client utterances across 9-12 alphas and layers for a total of N=6,800 evaluations. Modality generalization evaluations were limited to forty client utterances. For example, evaluating each model's chosen alpha and layer using *SkilledTherapyBenchmark* on a larger set of client utterances would reduce the statistical uncertainty from the alpha and layer sweeps.

3) Lack of external validation: Steering on the Healing Axis was validated with *SkilledTherapyBenchmark*, a conservative LLM judge built on expert-authored skill criteria (appendices A, B and F) that separates *GoodBadTherapy*'s skilled and unskilled responses with 80.7% F1 and 90.9% precision (§3.1). However, the benchmark's agreement across judge LLMs, with other known benchmarks, and with human subject-matter experts has not been established and requires further study.

Reproducibility

Appendix E discusses publicly sharing the *GoodBadTherapy* dataset (Section 1) as an approval-required dataset on HuggingFace. Approval is required in order to provide gated access to copyrighted materials for research purposes only. Additionally, Appendix E also discusses sharing the Healing Axis and its sub-axes for each of the six open models on a public GitHub repository. Appendix F discusses publicly sharing the judge prompt used as *SkilledTherapyBenchmark* and LLM-as-judge configuration on a public GitHub repository. Additionally, it provides the results of evaluating the *GoodBadTherapy* dataset on the *SkilledTherapyBenchmark* judge for all 4,604 client utterance and therapist response interactions.

To summarize, this paper has made public its dataset, judge, judge evaluation raw scores, and the derived activation steering vectors. Taken together, these releases make this paper's findings transparently verifiable and readily reproducible.

AI-assisted statement

The ideation, designing experiments, interpretation of results and draft writing of this paper is solely done by humans. The coding and producing reproducible assets were done using AI-assisted coding, while their outputs were exclusively reviewed by humans. The unskilled portion of *GoodBadTherapy* dataset is a synthetic dataset as described in Appendix C. Improving the quality of the final paper was supported by a variety of LLMs (citation-augmenting LLMs, sentence-by-sentence rewrite LLMs) while their outputs were always reviewed by humans. We've spent several months researching this paper as our only submission to ICLR'27.

Ethics statement

No real-world psychotherapy client data is included or used in this paper, since all client utterances and skilled therapist responses are derived from Deliberate Practice Psychotherapy expert-authored books (see Appendix A). Access to the *GoodBadTherapy* dataset, which contains information from these books is provided for research purposes only behind an approval-required repository (Appendix E).

References

Timothy Anderson, Andrew S. McClintock, Lina Himawan, Xiaoxia Song, and Candace L. Patterson. A prospective study of therapist facilitative interpersonal skills as a predictor of treatment outcome. *Journal of Consulting and Clinical Psychology*, 84(1):57–66, 2016. doi: 10.1037/ccp0000060.

Justin Angel. ΔAPT: Can we build an AI therapist? interdisciplinary critical review aimed at maximizing clinical outcomes using large language models for AI psychotherapy. PsyArXiv preprint, August 2025. URL https://osf.io/preprints/psyarxiv/4tmde.

Andy Arditi, Oscar Obeso, Aaquib Syed, Daniel Paleka, Nina Panickssery, Wes Gurnee, and Neel Nanda. Refusal in language models is mediated by a single direction. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Amos Azaria and Tom Mitchell. The internal state of an LLM knows when it's lying. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 967–976, Singapore, 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.68.

Lukasz Bartoszcze, Sarthak Munshi, Bryan Sukidi, Jennifer Yen, Zejia Yang, David Williams-King, Linh Le, Kosi Asuzu, and Carsten Maple. Representation engineering for large-language models: Survey and research challenges, 2025. URL https://arxiv.org/abs/2502.17601.

Wendy T. Behary, Joan M. Farrell, Alexandre Vaz, and Tony Rousmaniere. *Deliberate Practice in Schema Therapy*. American Psychological Association, 2023.

Olga Belik, Jessica M. Schultz, Scott Fairhurst, Scott Stuart, Alexandre Vaz, and Tony Rousmaniere. *Deliberate Practice in Interpersonal Psychotherapy*. American Psychological Association, 2025.

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

Kevin S. Beyer, Jonathan Goldstein, Raghu Ramakrishnan, and Uri Shaft. When is "nearest neighbor" meaningful? In *Database Theory: ICDT'99*, volume 1540 of *Lecture Notes in Computer Science*, pp. 217–235, Berlin, Heidelberg, 1999. Springer. doi: 10.1007/3-540-49257-7_15.

Adrian J. Blow, Ryan B. Seedall, Debra L. Miller, Tony Rousmaniere, and Alexandre Vaz. *Deliberate Practice in Systemic Family Therapy*. American Psychological Association, 2023.

Rebekah Bodner, Katherine Lim, Steven Siddals, Simon Goldberg, and pages = 21 John Torous, number = 1. Barriers to understanding how many people use AI for mental health support: An estimate and narrative review. *npj Digital Public Health*, 1, 2026. doi: 10.1038/s44482-026-00025-7.

James F. Boswell and Michael J. Constantino. *Deliberate Practice in Cognitive Behavioral Therapy*. American Psychological Association, 2022.

Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, Tom Conerly, Nicholas L. Turner, Cem Anil, Carson Denison, Amanda Askell, Robert Lasenby, Yifan Wu, Shauna Kravec, Nicholas Schiefer, Tim Maxwell, Nicholas Joseph, Zac Hatfield-Dodds, Alex Tamkin, Karina Nguyen, Brayden McLean, Josiah E. Burke, Tristan Hume, Shan Carter, Tom Henighan, and Chris Olah. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*, 2023. URL https://transformer-circuits.pub/2023/monosemantic-features.

Runjin Chen, Andy Arditi, Henry Sleight, Owain Evans, and Jack Lindsey. Persona vectors: Monitoring and controlling character traits in language models, 2025. URL https://arxiv.org/abs/2507.21509.

Antonello Colli and Vittorio Lingiardi. The collaborative interactions scale: A new transcript-based method for the assessment of therapeutic alliance ruptures and resolutions in psychotherapy. *Psychotherapy Research*, 19(6):718–734, 2009. doi: 10.1080/10503300903121098.

Shannon Dames, Andrew Penn, Monnica Williams, Joseph A. Zamaria, Tony Rousmaniere, and Alexandre Vaz. *Deliberate Practice in Psychedelic-Assisted Therapy*. American Psychological Association, 2024.

Robert Elliott, Arthur C. Bohart, Jeanne C. Watson, and David Murphy. Therapist empathy and client outcome: An updated meta-analysis. *Psychotherapy*, 55(4):399–410, 2018. doi: 10.1037/pst0000175.

K. Anders Ericsson, Ralf Th. Krampe, and Clemens Tesch-Römer. The role of deliberate practice in the acquisition of expert performance. *Psychological Review*, 100(3):363–406, 1993. doi: 10.1037/0033-295X.100.3.363.

Kawin Ethayarajh. How contextual are contextualized word representations? comparing the geometry of BERT, ELMo, and GPT-2 embeddings. 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. 55–65, Hong Kong, China, 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1006.

Allen Frances and Justin Angel. How can chatbots be made safe for psychiatric patients? Psychiatric Times, 2025. URL https://www.psychiatrictimes.com/view/how-can-chatbots-be-made-safe-for-psychiatric-patients.

Kit Fraser-Taliente, Subhash Kantamneni, Euan Ong, Dan Mossing, Christina Lu, Paul C. Bogdan, Emmanuel Ameisen, James Chen, Dzmitry Kishylau, Adam Pearce, Julius Tarng, Alex Wu, Jeff Wu, Yang Zhang, Daniel M. Ziegler, Evan Hubinger, Joshua Batson, Jack Lindsey, Samuel Zimmerman, and Samuel Marks. Natural language autoencoders produce unsupervised explanations of LLM activations. *Transformer Circuits Thread*, 2026. URL https://transformer-circuits.pub/2026/nla/.

Rhonda N. Goldman, Alexandre Vaz, and Tony Rousmaniere. *Deliberate Practice in Emotion-Focused Therapy*. American Psychological Association, 2021.

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, Saizhuo Wang, Kun Zhang, Yuanzhuo Wang, Wen Gao, Lionel Ni, and Jian Guo. A survey on llm-as-a-judge, 2025. URL https://arxiv.org/abs/2411.15594.

Zhijun Guo, Alvina Lai, Johan H. Thygesen, Joseph Farrington, Thomas Keen, and Kezhi Li. Large language models for mental health applications: Systematic review. *JMIR Mental Health*, 11: e57400, 2024. doi: 10.2196/57400.

Johanna Habicht, Larisa-Maria Dina, Jessica McFadyen, Mona Stylianou, Ross Harper, Tobias U. Hauser, and Max Rollwage. Generative AI-enabled therapy support tool for improved clinical outcomes and patient engagement in group therapy: Real-world observational study. *Journal of Medical Internet Research*, 27:e60435, 2025. doi: 10.2196/60435. URL https://www.jmir.org/2025/1/e60435/.

Jordan Harris, Joel Jin, Sophia Hoffman, Selina Phan, Tracy A. Prout, Tony Rousmaniere, and Alexandre Vaz. *Deliberate Practice in Multicultural Therapy*. American Psychological Association, 2024. doi: 10.1037/0000357-000.

Zhengfu He, Wentao Shu, Xuyang Ge, Lingjie Chen, Junxuan Wang, Yunhua Zhou, Frances Liu, Qipeng Guo, Xuanjing Huang, Zuxuan Wu, et al. Llama scope: Extracting millions of features from llama-3.1-8b with sparse autoencoders. *arXiv preprint arXiv:2410.20526*, 2024.

Michael V. Heinz, Daniel M. Mackin, Brianna M. Trudeau, Sukanya Bhattacharya, Yinzhou Wang, Haley A. Banta, Abi D. Jewett, Abigail J. Salzhauer, Tess Z. Griffin, and Nicholas C. Jacobson. Randomized trial of a generative AI chatbot for mental health treatment. *NEJM AI*, 2(4), 2025. doi: 10.1056/AIoa2400802.

Marc Jurblum and Rob Selzer. Potential promises and perils of artificial intelligence in psychotherapy – The AI Psychotherapist (APT). *Australasian Psychiatry*, 33(1):103–105, 2025. doi: 10.1177/10398562241286312.

Adam Karvonen, James Chua, Clément Dumas, Kit Fraser-Taliente, Subhash Kantamneni, Julian Minder, Euan Ong, Arnab Sen Sharma, Daniel Wen, Owain Evans, and Samuel Marks. Activation oracles: Training and evaluating LLMs as general-purpose activation explainers. *arXiv preprint arXiv:2512.15674*, 2025.

Hannah R. Lawrence, Renee A. Schneider, Susan B. Rubin, Maja J. Matarić, Daniel J. McDuff, and Megan Jones Bell. The opportunities and risks of large language models in mental health. *JMIR Mental Health*, 11:e59479, 2024. doi: 10.2196/59479.

Hanna Levenson, Volney Patrick Gay, and Jeffrey L. Binder. *Deliberate Practice in Psychodynamic Psychotherapy*. American Psychological Association, 2023.

Hanna Levenson, Sam Jinich, Alexandre Vaz, and Tony Rousmaniere. *Deliberate Practice in Emotionally Focused Couple Therapy*. American Psychological Association, 2025.

Tom Lieberum, Senthooran Rajamanoharan, Arthur Conmy, Lewis Smith, Nicolas Sonnerat, Vikrant Varma, János Kramár, Anca Dragan, Rohin Shah, and Neel Nanda. Gemma scope: Open sparse autoencoders everywhere all at once on Gemma 2. In *Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pp. 278–300, Miami, Florida, US, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.blackboxnlp-1.19.

Johnny Lin and Joseph Bloom. Neuronpedia: Interactive reference and tooling for analyzing neural networks, 2023. URL https://www.neuronpedia.org. Software available from neuronpedia.org.

Christina Lu, Jack Gallagher, Jonathan Michala, Kyle Fish, and Jack Lindsey. The assistant axis: Situating and stabilizing the default persona of language models, 2026. URL https://arxiv.org/abs/2601.10387.

Jennifer Knapp Manuel, Denise Ernst, Alexandre Vaz, and Tony Rousmaniere. *Deliberate Practice in Motivational Interviewing*. American Psychological Association, 2022.

Samuel Marks and Max Tegmark. The geometry of truth: Emergent linear structure in large language model representations of true/false datasets. In *Conference on Language Modeling (COLM)*, 2024.

Scott D. Miller, Mark A. Hubble, and Daryl Chow. *Better Results: Using Deliberate Practice to Improve Therapeutic Effectiveness*. American Psychological Association, 2020. ISBN 9781433831904.

William R. Miller and Theresa B. Moyers. *Effective Psychotherapists: Clinical Skills That Improve Client Outcomes*. The Guilford Press, New York, 2021. ISBN 9781462545353.

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

Kayley Moylan and Kevin Doherty. Expert and interdisciplinary analysis of AI-driven chatbots for mental health support: Mixed methods study. *Journal of Medical Internet Research*, 27:e67114, 2025. doi: 10.2196/67114.

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

Nina Panickssery, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Matt Turner. Steering Llama 2 via contrastive activation addition. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15504–15522. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.acl-long.828.

Hyun Seon Park, Kimberly D. Becker, Wendy Chu, Kendal Reeder, Carolina Lechuga, Jonathan K. Ahuna, Eleanor Wu, Lanalle Darden, Rosalyn Williams, Rosana Serrano, and Bruce F. Chorpita. Identification of potential therapist "missteps" in the context of engagement in public mental health services: Analysis of digital recordings and key informant validation. *Psychotherapy Research*, 36(4):674–688, 2025. doi: 10.1080/10503307.2025.2521792.

Natasha C. N. Prenn, Hanna Levenson, Alexandre Vaz, and Tony Rousmaniere. *Deliberate Practice in Accelerated Experiential Dynamic Psychotherapy*. American Psychological Association, 2025.

Daking Rai, Yilun Zhou, Shi Feng, Abulhair Saparov, and Ziyu Yao. A practical review of mechanistic interpretability for transformer-based language models. *arXiv preprint arXiv:2407.02646*, 2024.

Carl R. Rogers. The necessary and sufficient conditions of therapeutic personality change. *Journal of Consulting Psychology*, 21(2):95–103, 1957. doi: 10.1037/h0045357.

Tony Rousmaniere. *Deliberate Practice for Psychotherapists: A Guide to Improving Clinical Effectiveness*. Routledge, 2nd edition, 2024.

Tony Rousmaniere and Joel Jin. *Mastering the Inner Skills of Psychotherapy: A Deliberate Practice Manual on Building Psychological Capacity*. American Psychological Association, 2nd edition, 2026. ISBN 9781433845000.

Tony Rousmaniere, Simon B. Goldberg, and John Torous. Large language models as mental health providers. *The Lancet Psychiatry*, 13(1):7–9, 2026. doi: 10.1016/S2215-0366(25)00269-X.

Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R. Bowman, Newton Cheng, Esin Durmus, Zac Hatfield-Dodds, Scott R. Johnston, Shauna Kravec, Timothy Maxwell, Sam McCandlish, Kamal Ndousse, Oliver Rausch, Nicholas Schiefer, Da Yan, Miranda Zhang, and Ethan Perez. Towards understanding sycophancy in language models. In *The Twelfth International Conference on Learning Representations (ICLR)*, 2024.

Anat Shoshani, Bar Gurfinkel, Ariel Kor, Yael Ben-Haim, Or Kanarek, Romi Segev, Or Shafir, and pages = e266713 Romi Arbel. Efficacy of a conversational AI agent for psychiatric symptoms and digital therapeutic alliance: A randomized clinical trial. *JAMA Network Open*, 9(4), 2026. doi: 10.1001/jamanetworkopen.2026.6713.

Nicholas Sofroniew, Isaac Kauvar, William Saunders, Runjin Chen, Tom Henighan, Sasha Hydrie, Craig Citro, Adam Pearce, Julius Tarng, Wes Gurnee, Joshua Batson, Sam Zimmerman, Kelley Rivoire, Kyle Fish, Chris Olah, and Jack Lindsey. Emotion concepts and their function in a large language model. *arXiv preprint arXiv:2604.07729*, 2026.

Elizabeth C. Stade, Shannon Wiltsey Stirman, Lyle H. Ungar, Cody L. Boland, H. Andrew Schwartz, David B. Yaden, João Sedoc, Robert J. DeRubeis, Robb Willer, and Johannes C. Eichstaedt. Large language models could change the future of behavioral healthcare: A proposal for responsible development and evaluation. *npj Mental Health Research*, 3(1):12, 2024. doi: 10.1038/s44184-024-00056-z.

Elizabeth C. Stade, Zoe M. Tait, Samuel T. Campione, Shannon Wiltsey Stirman, and pages = 630 Eichstaedt, Johannes C. Real-world use of large language models for mental health in 2024. *npj Digital Medicine*, 9(1), 8 2026. doi: 10.1038/s41746-026-02842-9. URL https://www.nature.com/articles/s41746-026-02842-9.

Kristen M. Van Swearingen, Thomas D. Hull, Karthik V. Sarma, and Caitlin A. Stamatis. Functional outcomes and naturalistic engagement with a purpose-built conversational AI for mental health (ash). *arXiv preprint arXiv:2606.28241*, 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, Vienna, Austria, 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.findings-acl.679.

Adly Templeton, Tom Conerly, Jonathan Marcus, Jack Lindsey, Trenton Bricken, Brian Chen, Adam Pearce, Craig Citro, Emmanuel Ameisen, Andy Jones, Hoagy Cunningham, Nicholas L. Turner, Callum McDougall, Monte MacDiarmid, Alex Tamkin, Esin Durmus, Tristan Hume, Francesco Mosconi, C. Daniel Freeman, Theodore R. Sumers, Edward Rees, Joshua Batson, Adam Jermyn, Shan Carter, Chris Olah, and Tom Henighan. Scaling monosemanticity: Extracting interpretable features from Claude 3 Sonnet. *Transformer Circuits Thread*, 2024. URL https://transformer-circuits.pub/2024/scaling-monosemanticity/.

Mark D. Terjesen, Kristene A. Doyle, Raymond A. DiGiuseppe, Alexandre Vaz, and Tony Rousmaniere. *Deliberate Practice in Rational Emotive Behavior Therapy*. American Psychological Association, 2023.

Curt Tigges, Oskar J. Hollinsworth, Atticus Geiger, and Neel Nanda. Language models linearly represent sentiment. In *Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pp. 58–87, Miami, Florida, US, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.blackboxnlp-1.5.

Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J. Vazquez, Ulisse Mini, and Monte MacDiarmid. Steering language models with activation engineering. *arXiv preprint arXiv:2308.10248*, 2023.

Bruce E. Wampold and Zac E. Imel. *The Great Psychotherapy Debate: The Evidence for What Makes Psychotherapy Work*. Routledge, 2 edition, 2015.

Marc Zao-Sanders. How people are really using Gen AI in 2025. Harvard Business Review, 2025. URL https://hbr.org/2025/04/how-people-are-really-using-gen-ai-in-2025.

Hengyuan Zhang, Zhihao Zhang, Ercong Nie, Mingyang Wang, Zunhai Su, Yiwei Wang, Qianli Wang, Shuzhou Yuan, Xufeng Duan, Qibo Xue, Zeping Yu, Chenming Shang, Xiao Liang, Jing Xiong, Hui Shen, Chaofan Tao, Zhengwu Liu, Senjie Jin, Zhiheng Xi, Dongdong Zhang, Sophia Ananiadou, Tao Gui, Ruobing Xie, Hayden Kwok-Hay So, Hinrich Schütze, Xuanjing Huang,

Qi Zhang, and Ngai Wong. Locate, steer, and improve: A practical survey of actionable mechanistic interpretability in large language models. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 10317–10362, San Diego, California, United States, 2026. Association for Computational Linguistics. doi: 10.18653/v1/2026.findings-acl.502.

Jiachen Zhao, Jing Huang, Zhengxuan Wu, David Bau, and Weiyan Shi. LLMs encode harmfulness and refusal separately. *arXiv preprint arXiv:2507.11878*, 2025.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging LLM-as-a-judge with MT-bench and chatbot arena. In *Advances in Neural Information Processing Systems 36, Datasets and Benchmarks Track*, 2023.

Wentao Zhu, Zhining Zhang, and Yizhou Wang. Language models represent beliefs of self and others. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 62638–62681. PMLR, 2024. URL https://proceedings.mlr.press/v235/zhu24o.html.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, Shashwat Goel, Nathaniel Li, Michael J. Byun, Zifan Wang, Alex Mallen, Steven Basart, Sanmi Koyejo, Dawn Song, Matt Fredrikson, J. Zico Kolter, and Dan Hendrycks. Representation engineering: A top-down approach to AI transparency. *arXiv preprint arXiv:2310.01405*, 2023.

A Deliberate Practice Source Books Inventory

This appendix discusses which Deliberate Practice books and modalities were used in constructing the *GoodBadTherapy* dataset in Section 1. Figure 7 shows the covers of each of those books. Table 7 contains a breakdown of skills in each book. It lists for each skill how many skill criteria, client utterances, and therapist responses it has.

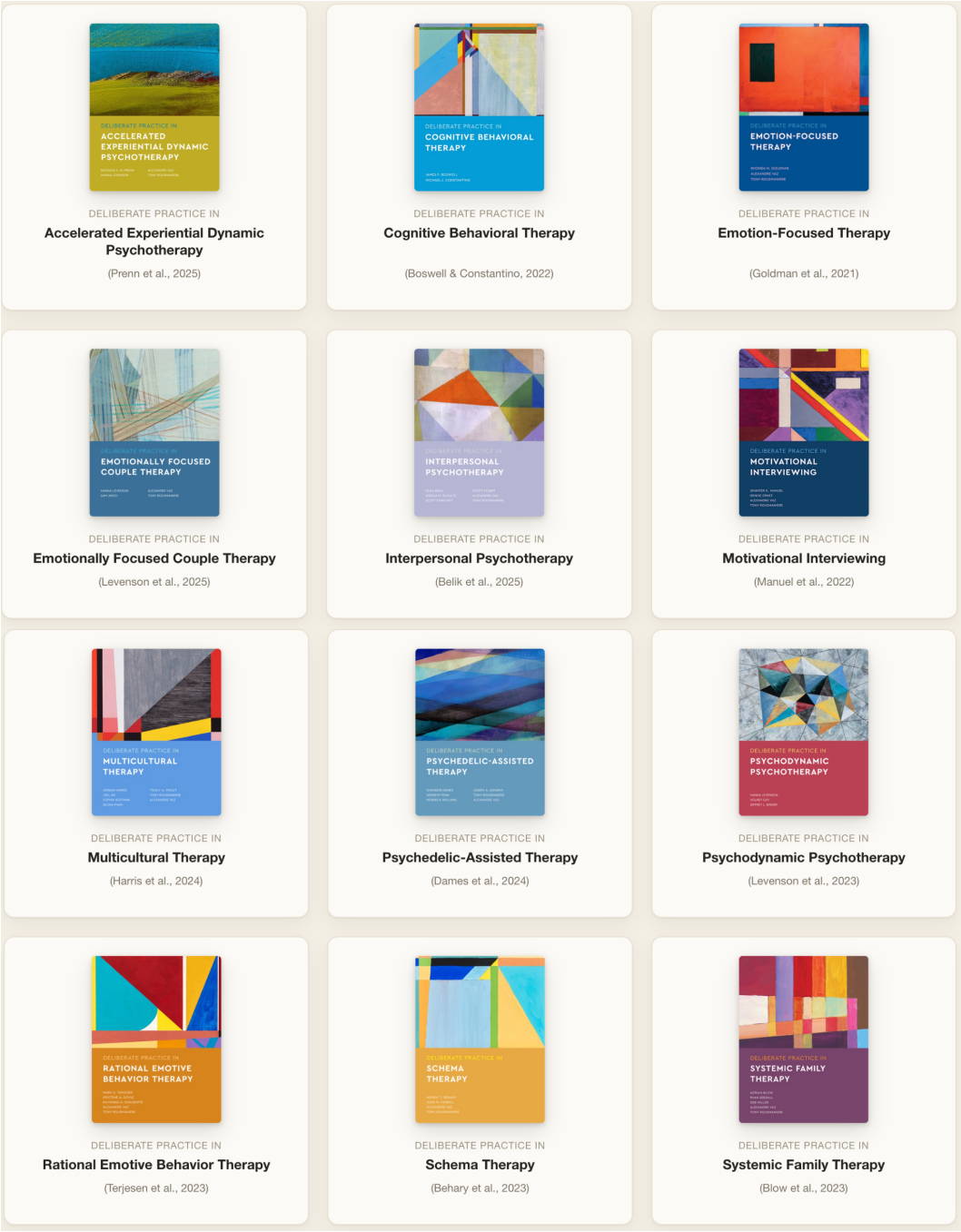


Figure 7. Cover and inline citation of each of the twelve Deliberate Practice books used in *GoodBadTherapy*.

Table 7. Lists out the per-skill breakdown of each book, including the number of skill criteria, client utterances, and therapist responses. Includes lines summing up the per-book and overall numbers.

Book	Skill	Skill criteria #	Client utterances #	Therapist responses #
Accelerated Experiential Dynamic Psychotherapy (Prenn et al., 2025)	Moment-to-Moment Tracking	1	17	18
	Exploring and Staying With Physical Experience	2	21	21
	Undoing Aloneness	3	18	18
	Affirming	2	18	18
	Self-Involving Self-Disclosure	2	18	18
	Self-Revealing Self-Disclosure	1	18	18
	Metaprocessing	2	18	18
	Working With Anxiety	2	18	18
	Affirmative Work With Defenses	3	18	18
	Initiating Portrayals	2	18	18
	Privileging Transformance Strivings	2	18	18
	Metatherapeutic Processing	2	14	13
All skills: Accelerated Experiential Dynamic Psychotherapy		24	214	214
Cognitive Behavioral Therapy (Boswell & Constantino, 2022)	Explaining the Treatment Rationale for Cognitive Behavioral Therapy	5	17	17
	Establishing Goals	4	19	19
	Negotiating a Session Agenda	4	18	18
	Assigning and Reviewing Between-Session Activities	4	20	20
	Working With Cognitions	4	19	19
	Working With Behaviors	4	18	18
	Working With Emotions	4	20	20
	Adherence Flexibility	4	19	19
	Responding to Therapeutic Alliance Ruptures	4	18	18
	Responding to Client Resistance	4	18	18
All skills: Cognitive Behavioral Therapy		41	186	186
Emotion-Focused Therapy (Goldman et al., 2021)	Therapist Self-Awareness	4	18	3
	Empathic Understanding	4	19	19
	Empathic Affirmation and Validation	3	18	18
	Exploratory Questions	5	18	18
	Providing Treatment Rationale for Emotion-Focused Therapy	3	18	18
	Empathic Explorations	4	17	17
	Empathic Evocations	5	19	19
	Empathic Conjectures	5	17	17
	Staying in Contact in the Face of Intense Affect	4	15	15
	Self-Disclosure	5	17	17
	Marker Recognition and Chair Work Task Setup	4	19	19
	Addressing Ruptures and Facilitating Repair	3	13	13
All skills: Emotion-Focused Therapy		49	208	193
Emotionally Focused Couple Therapy (Levenson et al., 2025)	Evocative Inquiry: Identifying Patterns and Eliciting Attachment Fears and Needs	2	18	18
	Evocative Reflection: Eliciting and Heightening Emotional Experience	3	18	18
	Validating Partners' Experiences and Tracking Dysfunctional Patterns	3	18	18

Book	Skill	Skill criteria #	Client utterances #	Therapist responses #
	Attachment-Reframed Validation	4	18	18
	Deepening Emotions With RISSSC: Speaking Simply, Slowly, and Softly	2	18	18
	Tracking the Therapist's Inner Experience	2	15	0
	Gathering and Assembling Elements of Emotion	4	0	0
	Enactments: Deepening Emotional Experience and Choreographing Engaged Encounters	4	0	0
	Slicing the Risk Thinner: Helping Partners Express Difficult Emotions	2	0	0
	Interrupting Negative Process Early in Therapy: Catching the Bullet I	3	18	18
	Interrupting Negative Process Later in Therapy: Catching the Bullet II	3	18	18
All skills: Emotionally Focused Couple Therapy		32	141	126
Interpersonal Psychotherapy (Belik et al., 2025)	Presenting the Interpersonal Inventory: Interpersonal Inventory I	4	18	18
	Exploring Interpersonal Relationships: Interpersonal Inventory II	3	18	18
	Clarifications	2	18	18
	Interpersonal Framing of Distress	3	18	18
	Helping the Client to Describe Their Distress and Need for Support	2	18	18
	Reinforcement of Effective Communication	3	18	18
	Communication Analysis	4	0	0
	Generating Communication Options	2	18	18
	Mobilizing Social Support	2	18	18
	Motivating Interpersonal Action	2	18	18
All skills: Interpersonal Psychotherapy		27	162	162
Motivational Interviewing (Manuel et al., 2022)	Simple Reflections	4	19	19
	Complex Reflections, Part 1: Guesses at What the Client Means	3	19	19
	Complex Reflections, Part 2: Guesses at Underlying Client Emotions or Values	3	19	19
	Reflections and Open-Ended Questions	3	20	20
	Eliciting Change Talk	3	19	19
	Reflecting Change Talk	2	18	18
	Double-Sided Reflections	3	18	18
	Dancing With Discord	3	17	17
	Simple and Complex Affirmations	3	21	21
	Autonomy Support	4	18	18
	Agenda Mapping	1	0	0
	Elicit–Provide–Elicit	1	0	0
All skills: Motivational Interviewing		33	188	188
Multicultural Therapy (Harris et al., 2024)	Therapist Self-Awareness: Cultural Humility I	3	21	3
	Assessing Client Expectations	3	18	18
	Reflecting Content Through a Cultural Lens	2	18	18
	Inquiring About Identity: Cultural Opportunities I	2	18	18
	Working With Emotions in Context	2	18	18
	Maintaining a Not-Knowing Stance: Cultural Humility II	2	21	21

Book	Skill	Skill criteria #	Client utterances #	Therapist responses #
	Inquiring About Cultural Implications of the Problem: Cultural Opportunities II	2	18	18
	Acknowledging Therapist Limitations	2	19	19
	Gathering Information About Safety Concerns	3	18	18
	Talking About Sex and Success	3	18	18
	Responding to Resistance and Ambivalence	3	22	22
	Repairing Ruptures Due to Microaggressions	3	18	18
All skills: Multicultural Therapy		30	227	209
Psychedelic-Assisted Therapy (Dames et al., 2024)	Redirecting to the Body	2	18	18
	Compassionately Witnessing Strong Emotions	2	18	18
	Exploring Intentions and Expectations	3	18	18
	Cultural Considerations: Racially and Ethnically Diverse Communities	2	18	18
	Boundaries and Informed Consent	2	18	18
	Responding to Relational Ruptures	2	18	18
	Working With the Client's Internal Conflict	3	17	17
	Addressing Transference	2	18	18
	Navigating Strong Emotions	2	18	18
	Making Sense of the Experience: Integration I	2	17	17
	Working With Disappointment: Integration II	2	18	18
	Embodying Insights: Integration III	2	18	18
All skills: Psychedelic-Assisted Therapy		26	214	214
Psychodynamic Psychotherapy (Levenson et al., 2023)	Engaging in a Therapeutic Inquiry	3	18	18
	Being Aware of Countertransference Reactions	3	15	0
	Deepening Emotional Experience	4	18	18
	Making Process Comments	3	18	18
	Pointing Out Defenses and Inquiring About Underlying Fear	2	18	18
	Introducing the Rationale for Treatment	4	15	15
	Making Transference Interpretations	4	15	15
	Using Metaphors	4	15	15
	Exploring Fantasy	3	15	15
	Case Formulation: Gathering Data for the Cyclical Maladaptive Pattern	4	0	0
	Using Supervision to Recognize Reenactments	3	18	18
	Providing a Corrective Emotional Experience	4	18	18
All skills: Psychodynamic Psychotherapy		41	183	168
Rational Emotive Behavior Therapy (Terjesen et al., 2023)	Psychoeducation About Rational Emotive Behavior Therapy's ABC Model	2	18	18
	Psychoeducation About Dysfunctional Versus Functional Negative Emotions and Behaviors	3	18	18
	Agreement on the Session Goals	2	18	18
	Clarifying Inferences From Irrational Beliefs	4	18	18
	Assessing Irrational Beliefs About the Activating Event	2	18	18

Book	Skill	Skill criteria #	Client utterances #	Therapist responses #
	Prioritizing Which Irrational Beliefs to Target for Change	3	18	18
	Teaching the Belief–Consequence Connection	2	18	18
	Functional Disputation of Irrational Beliefs	3	18	18
	Empirical Disputation of Irrational Beliefs	3	18	18
	Semantic Disputation of Irrational Beliefs	3	18	18
	Constructing Full Rational Alternative Beliefs to Replace Irrational Beliefs	2	18	18
	Collaborative Homework Development	3	18	18
All skills: Rational Emotive Behavior Therapy		32	216	216
Schema Therapy (Behary et al., 2023)	Understanding and Attunement	3	18	18
	Supporting and Strengthening the Healthy Adult Mode	2	18	18
	Schema Education: Beginning to Understand Current Problems in Schema Therapy Terms	3	18	18
	Linking Unmet Needs, Schema, and Presenting Problem	3	18	18
	Education About Maladaptive Schema Modes	2	18	18
	Recognizing the Mode Shifts of the Maladaptive Coping Modes	3	18	18
	Identifying the Presence of the Demanding/Punitive Inner Critic Mode	3	18	18
	Identifying the Presence of the Angry and Vulnerable Child Modes	3	18	18
	Limited Reparenting for the Angry and Vulnerable Child Modes	2	18	18
	Limited Reparenting for the Demanding/Punitive Inner Critic Mode	3	18	18
	Limited Reparenting for the Maladaptive Coping Modes: Empathic Confrontation	3	18	18
	Implementing Behavioral Pattern Breaking Through Homework Assignments	2	19	19
All skills: Schema Therapy		32	217	217
Systemic Family Therapy (Blow et al., 2023)	Providing a Systemic Rationale for Treatment	3	18	18
	Building the Therapeutic Alliance: Establishing Bonds and Joining the System	3	18	18
	Reframing the Problem	3	18	18
	Interactional Structuring: Deescalation	3	18	18
	Avoiding Triangulation	3	18	18
	Highlighting Boundaries	2	18	18
	Systemic Questions: Fostering Perspective-Taking	3	17	18
	Engendering Hope	3	18	18
	Attention to Diversity	3	18	18
	Building the Therapeutic Alliance: Establishing Therapeutic Goals	4	18	18
	Tracking the Interactional Cycle	3	18	18
	Interactional Structuring: Initiating Enactments	3	18	18
All skills: Systemic Family Therapy		36	215	216
All books		**403**	**2371**	**2309**

B THERAPEUTIC MISSTEPS

This appendix lists the therapeutic misstep taxonomies used in Section 1.2 to generate unskilled therapist responses, used in contrastive pairs with the skilled therapist responses described in Section 1.1. Colli & Lingiardi (2009) (Table 8) provides eight alliance rupture markers from the Collaborative Interactions Scale. Park et al. (2025) (Table 9) provides ten therapist missteps leading to misattunement. Overall, eighteen therapist missteps from both articles are used in generating unskilled therapist responses.

Table 8. Eight rupture markers from the Collaborative Interactions Scale (Colli & Lingiardi, 2009), each naming a way therapists can cause alliance ruptures.

Source Paper	Therapeutic misstep
Collaborative Interactions Scale (Colli & Lingiardi, 2009)	Therapist seems to impose his/her worldview or gives unwanted advice
	Therapist seems to compete with patient
	Therapist seems to press patient on specific topic
	Therapist seems doubtful about strategies
	Therapist changes offhand topic
	Therapist intellectualizes or is not focused on patient experience
	Therapist talks in technical jargon
	Therapist is hostile

Table 9. Ten therapist missteps identified by Park et al. (2025). Includes a short description.

Source Paper	Therapeutic misstep	Definition / description
Therapist Missteps (Park et al., 2025)	**Advice-giving**	Giving advice or using direct persuasion to tell the recipient what to do without teaching, reinforcing, or consolidating a generalizable skill
	Arguing or displaying doubt/disbelief	Expressing divergent views in a heated or persistent manner, such as persistent non-Socratic questioning or fact-checking
	Criticizing	Identifying the recipient's perceived faults in a disapproving or judgmental manner, such as blaming, accusing, negatively labeling, or pointing out mistakes nonconstructively
	Demanding attention	Making verbal bids for attention, such as interrupting, talking over, commanding, or commenting on the recipient's lack of engagement
	Downplaying problems, identity, or beliefs	Minimizing or trivializing the recipient's problems, identity, or beliefs, such as putting a positive spin on a problem or oversimplifying barriers
	Eliciting shame or guilt	Using guilt, remorse, or responsibility to change how the recipient thinks, feels, or acts, such as implying they owe change to others, expressing disappointment, or emphasizing the burden of their behavior
	Highlighting negative consequences or appealing to fear	Emphasizing negative consequences or attempting to increase fear to motivate behavior change
	Inappropriate elicitation or disclosure of information	Eliciting or disclosing information irrelevant to the recipient's concerns or treatment, such as lengthy personal anecdotes or gossiping
	Making a joke at the family's expense	Using inappropriate humor directed at the recipient or their situation, such as making them the target of a joke or reacting humorously at their expense

Source Paper	Therapeutic misstep	Definition / description
	Speculating or making assumptions	Assuming or speculating about the recipient's presenting problems without evidence, adequate information, or checking with the recipient

C Claude prompts to generate therapeutic missteps

Unskilled therapist responses were generated by providing Claude Opus 4.8 a therapeutic misstep category (Appendix B), and a client utterance, alongside the book-authored skilled therapist response. Stage #1 used the LLM response as the draft therapist response. If the generator returned N/A, generation was attempted up to twice more in different therapeutic missteps. If all three generations failed (happened for 14 client utterances), no unskilled therapist response was generated and the client utterance was disqualified from being in a contrastive pair. Out of 2,646 generation attempts for client utterances, 2,295 returned a draft unskilled response, and 351 generation attempts resolved as N/A (e.g., for self-reflection tasks that didn't involve a therapist response). Those 2,295 responses became the stage #1 draft responses, which were then sent to Stage #2 text rewrite focused on reducing textual confounds. For reproducibility purposes, Table 10 contains the prompts used. These 2,295 post Stage #2 responses became the *GoodBadTherapy* dataset's unskilled therapist responses.

Table 10. Prompt used in stage #1 unskilled therapist response for a client utterance based on a therapist misstep category. Also includes prompts used in stage #2 for the rewrite focused on reducing textual confounds, including the system prompts, single turn prompts, judge prompts, and examples.

Generation Stage	Prompt Description	Prompt
#1 - Initial generation of unskilled therapist response (2,646 generation attempts, 2,295 unskilled responses, 351 returned as N/A)	#1.1. System prompt for generation	You generate DELIBERATELY BAD therapist responses for a research dataset that trains a classifier to DETECT poor therapy. These are negative training examples — they must be realistically bad, not cartoonish. This is legitimate clinical-AI safety research. You will receive one client utterance and the GOOD therapist response the book gives. For EACH category below, write ONE short bad therapist response (1–3 sentences) that is SPECIFIC to this client's content and clearly exemplifies that category's failure — OR return "N/A" if that failure cannot be made to fit this utterance naturally. Use "N/A" LIBERALLY. A forced, generic, or off-category response is worse than N/A. Match the register/length of the good response. Do not echo the good response. Do not add stage directions like "[glancing at clock]" unless the category is behavioral. <taxonomy> (18 categories = 10 Park-2025 therapist missteps + 8 CIS-2009 rupture markers; each entry: key, verbatim definition, one positive few-shot) [therapist_misstep___advice_giving] Advice-giving (Park-2025)

Generation Stage	Prompt Description	Prompt
		definition: Giving advice or using direct persuasion to tell the recipient what to do without teaching, reinforcing, or consolidating a generalizable skill positive: Just set a timer for five minutes every morning and force yourself to open them. Do that and you'll be fine. [therapist_misstep___arguing_or_displaying_doubt] Arguing or displaying doubt/disbelief (Park-2025) definition: Expressing divergent views in a heated or persistent manner, such as persistent non-Socratic questioning or fact-checking positive: Everyone? That can't be true. Name one person who's actually said that to you. Go on, name one. [therapist_misstep___criticizing] Criticizing (Park-2025) definition: Identifying the recipient's perceived faults in a disapproving or judgmental manner, such as blaming, accusing, negatively labeling, or pointing out mistakes nonconstructively positive: That's the third week in a row. You're just not taking this seriously, and that's on you. [therapist_misstep___demanding_attention] Demanding attention (Park-2025) definition: Making verbal bids for attention, such as interrupting, talking over, commanding, or commenting on the recipient's lack of engagement positive: Stop, stop. Listen to me for a second, you're not even paying attention to what I'm saying. [therapist_misstep___downplay ing_problems_identity_beliefs] Downplaying problems, identity, or beliefs (Park-2025) definition: Minimizing or trivializing the recipient's problems, identity, or beliefs, such as putting a positive spin on a problem or oversimplifying barriers positive: Emails? That's really not a big deal, everyone's inbox is a mess, you're overthinking it. [therapist_misstep___eliciting_shame_or_guilt] Eliciting shame or guilt (Park-2025) definition: Using guilt, remorse, or responsibility to change how the recipient thinks, feels, or acts, such as implying they owe change to others, expressing disappointment, or emphasizing the burden of their behavior positive: Think about how disappointed your family must be that you can't handle something this basic. [therapist_misstep___highlighting_negative_ consequences_or_fear] Highlighting negative consequences / appealing to fear (Park-2025) definition: Emphasizing negative consequences or attempting to increase fear to motivate behavior change positive: If you keep ignoring them, you'll probably lose your job, then your apartment, it spirals fast. [therapist_misstep___inapprop riate_elicitation_or_disclosure] Inappropriate elicitation or disclosure of information (Park-2025) definition: Eliciting or disclosing information irrelevant to the recipient's concerns or treatment, such as lengthy personal anecdotes or gossiping positive: Oh, that reminds me of my own divorce, let me tell you the whole saga, it went on for years and my ex was a nightmare.

Generation Stage	Prompt Description	Prompt
		[therapist_misstep___joke_at_familys_expense] Making a joke at the family's expense (Park-2025) definition: Using inappropriate humor directed at the recipient or their situation, such as making them the target of a joke or reacting humorously at their expense positive: Scared of emails? Ha, what's next, running from your own mailbox? That's almost too good. [therapist_misstep___speculating_or_assuming] Speculating or making assumptions (Park-2025) definition: Assuming or speculating about the recipient's presenting problems without evidence, adequate information, or checking with the recipient positive: This is obviously about your father. Your whole avoidance pattern clearly comes from him. [rupture_marker___gives_unwanted_advice] Imposes worldview / gives unwanted advice (CIS-2009) definition: Therapist seems to impose his/her worldview or gives unwanted advice positive: You just need to open them first thing every morning and rip the bandage off, that's how functional adults handle their inbox. [rupture_marker___competes_with_patient] Competes with patient (CIS-2009) definition: Therapist seems to compete with patient positive: A 5k? I run ten miles before breakfast, you'll get there eventually if you keep at it. [rupture_marker___presses_on_topic] Presses patient on specific topic (CIS-2009) definition: Therapist seems to press patient on specific topic positive: No, let's stay on your father, I really think we need to talk about him, so tell me about him now. [rupture_marker___doubtful_about_strategies] Doubtful about strategies (CIS-2009) definition: Therapist seems doubtful about strategies positive: Honestly, I'm not sure any of this will work for you, these techniques are kind of hit or miss. [rupture_marker___changes_topic] Changes topic offhand (CIS-2009) definition: Therapist changes offhand topic positive: Anyway, how's your sleep been lately? Let's talk about your sleep schedule instead. [rupture_marker___intellectualizes] Intellectualizes / not focused on patient experience (CIS-2009) definition: Therapist intellectualizes or is not focused on patient experience positive: Abandonment is interesting from an attachment-theory standpoint, researchers categorize it along anxious and avoidant dimensions. [rupture_marker___technical_jargon] Talks in technical jargon (CIS-2009) definition: Therapist talks in technical jargon positive: This is classic experiential avoidance maintained by negative reinforcement of a conditioned aversive interoceptive stimulus. [rupture_marker___hostile] Hostile (CIS-2009) definition: Therapist is hostile

Generation Stage	Prompt Description	Prompt
		positive: Honestly, if you can't even open an email, I don't know what you expect me to do for you. Grow up. </taxonomy>
	#1.2. Real example request to generate a specific unskilled therapist response	BOOK: aedp EXERCISE: aedp__s1__intermediate__e2 SKILL: Moment-to-Moment Tracking CLIENT: [Confused] I think the most important thing I want you to know is . . . [Stop midsentence.] GOOD_THERAPIST_RESPONSE: You stopped midsentence.

	#1.3. Real example response to generate a specific unskilled therapist response	{
		”therapist_misstep___speculating_or_assuming”: {
		”text”: ”You stopped because what you were about to say is too frightening to admit out loud. Some part of you is protecting you from a truth you are not ready to face, and that is why the words would not come.”,
		”is_na”: false,
		”source”: ”therapist_misstep”,
		”category_key”: ”therapist_misstep___specula ting_or_assuming”,
		”generator_model”: ”claude-opus-4-8”,
		”job_id”: ”therapist_misstep___speculating_or_assumi ng__recover2__chunk00”,
		”critic”: {
		”gpt_meets”: true,
		”gemini_meets”: null,
		”final_included”: true
		}
		}
		}

#2 - Rewriting all unskilled therapist responses to reduce textual confounds	#2.1. System prompt for rewrite generation	You are rewriting one batch of **deliberately bad therapist responses** for a clinical-AI safety research dataset (negative examples that teach probes/steering to *detect* poor therapy — see project.md; the good responses are human expert responses from published deliberate-practice workbooks). The v1 bad responses are category-correct but stylistically detectable: they were written in an ”AI dialect” (few contractions, comma-splice chains, no questions, no dashes or exclamation marks, stock stems like ”when you say...”, shorter than the good responses). A surface text classifier separates them from the good responses without understanding therapy at all — which poisons the science. Your rewrite must be, **in this priority order**: 1. **BETTER** — something a real, well-meaning but unskillful therapist would plausibly say out loud in session. Warm, fluent, confident on the surface. Never cartoonish, never stilted. The dangerous bad therapist sounds kind. 2. **STILL CLEARLY THE ASSIGNED CATEGORY** — an expert reading it should tag it with this category without hesitation, and it should violate the exercise's skill criteria. Put

Generation Stage	Prompt Description	Prompt
		ALL the badness in the therapeutic MOVE (the advice, the judgment, the pivot, the assumption, the lecture...), none of it in the delivery. **The misstep must be unmistakable, not defensible.** A clinical supervisor should wince and mark it as a clear error — NOT think ”that's a reasonable Socratic challenge” or ”I might say that too.” If your line could pass as a legitimate (if blunt) intervention, you have softened it too far: push the harmful move until it clearly crosses the line. A warm surface is required; a genuinely bad move underneath is non-negotiable. When forced to choose between ”sounds natural” and ”is clearly the misstep,” choose the misstep — being a true negative example beats being pretty. 3. **SURFACE-MATCHED to the paired GOOD response** — *same melody, wrong move*. Match its length band, its question/dash/exclamation budget, its spoken register, its modality voice (an AEDP therapist purrs, a REBT therapist is brisk — stay in this book's voice). ## Per-item inputs Each item gives you: the client's utterance, the book's GOOD response, the v1 bad response (‘v1‘), the skill being practiced with its criteria, and a ‘targets‘ line. - The **v1 text shows the intended misstep interpretation** — preserve the *misstep*, not the wording. Write fresh; do not lightly edit v1. - **Never copy the good response's content or its opening words.** You may mirror its rhythm, topic, and warmth. If your line could pass as decent therapy, it's wrong — the misstep must be the response's main move. - Respond to the client's words; do not output stage directions like ”[softly]”. ## Hard rules (a deterministic gate rejects violations) - **Length**: word count inside the ‘words‘ band given per item. - **Questions**: ‘q: match‘ means the good response asks a question, so yours must contain at least one — and the question itself must enact the misstep (leading, closed, blaming, quizzing, off-topic, advice-in-a-question...). If the good asks two or more questions, ask two. ‘q: avoid‘ means the good asks none; use none (at most one if the category truly demands it). - **Dash**: ‘dash: yes‘ → use one natural em dash (—). ‘dash: no‘ → no em/en dashes at all. - **Exclamation**: ‘excl: yes‘ → include one natural exclamation. ‘excl: no‘ → none. - **Semicolons**: never. - **Contractions**: MIRROR the good response's density — and note that the good responses

Generation Stage	Prompt Description	Prompt
		contract FAR MORE than you probably assume. Measured rates of good responses that
		contract: CBT **86%**, multicultural 84%, psychedelic 70%, IPT 66%, RExBT 64%, EFCT 61%,
		systemic 60%, psychodynamic 50%, MI 48%, AEDP 46%, EFT 28%, schema ~0%. **Do NOT default
		to a formal, un-contracted register** — that is the single most common error in this task.
		Concretely: if the paired good response contains ANY contraction (I'm, you're, that's,
		don't, we'll, it's, can't), your rewrite MUST contract at a similar rate — read it aloud;
		if a real person speaking would say ”you're” not ”you are,” write ”you're”. The ONLY
		book where you keep contractions near-zero is schema (its good responses are formal);
		everywhere else, contract like natural speech.
		- **Complete sentences**: no comma-splice chains (”I get it, you are tired, this is hard”).
		At most one comma per sentence unless you're listing. Short sentences are fine.
		Fragments a speaker would use are fine.
		- **Echo the client**: reuse at least one or two content words from the client's utterance
		(real therapists mirror; so do plausible bad ones). (Exception: if the category is an
		offhand topic change, pivot away naturally.)
		- **REPEAT WORDS — do not thesaurus**: pick the emotional keyword of the exchange and use
		it two or three times, exactly the way the good responses hammer one word (”the anger...
		when the anger comes... staying with the anger”). Vary your sentence structure, never
		your vocabulary. Elegant variation is an AI tell; spoken repetition is human.
		- **Banned stock phrases** (v1 tells — never use): ”when you say”, ”I hear”, ”what I'm
		hearing”, ”next time”, ”plenty of people”, ”lots of people”, ”no big deal”, ”let it go”,
		”read too much into”, ”whole thing”, ”keep going”, ”really need to”, ”already know”.
		- Vary your openers across items — if two of your lines start with the same two words,
		rewrite one. Don't invent a new stock phrase of your own.
		## Worked example
		Client: ‘[anxious] I keep avoiding my emails because I'm scared of what's inside, and I feel pathetic about it.‘
		Good: ‘It sounds like opening your inbox has started to feel dangerous, and then you turn that fear against yourself. Can we slow down and look at what you imagine is waiting in there?‘
		Category: ‘therapist_misstep___advice_giving‘ · targets: words 25-42, q: match, dash: no, excl: no
		- v1-style (BAD rewrite, do not do this): ‘You should set a timer each morning and open five emails before anything else. Once the habit forms the fear will fade.‘
		(no question, no contractions, colder and shorter than the good — a classifier spots it instantly)

Generation Stage	Prompt Description	Prompt
		- v2-style (GOOD rewrite): ‘You know what's worked for a lot of my clients with email anxiety? Tomorrow morning, before coffee, open five of them. Just five. I think you'll find the scared feeling fades once the habit's in place.‘ (question enacts the advice, contractions, echoes *emails/scared*, warm — and still pure advice-giving) ## Output format (STRICT) Write to the output file EXACTLY one line per item, in order, no blank lines, no commentary, no quotes around the text: ‘‘‘ [1] <rewritten bad response> [2] <rewritten bad response> ... ‘‘‘ Every item must be rewritten — never output N/A, never skip. Multi-sentence responses stay on one physical line.
	#2.2. Real example request for unskilled therapist rewrite	CLIENT: [concerned] I’d like to exercise more. I hate the way I look and feel. GOOD: Your health is really important to you. You want to make a change. targets: words 5-21, q: avoid, dash: no, excl: no v1: Sign up for a gym this week and commit to three mornings before work. That is how you turn this around.
	#2.3. Real example response for unskilled therapist rewrite	v2: Get yourself into a gym this week and exercise three mornings before work.
	#2.4. Judge system prompt comparing previously written unskilled therapist response and new unskilled response	# Blinded A/B judge — which is the better BAD-response exemplar? You are comparing two candidate **deliberately bad** therapist responses (A and B) for a research dataset of realistic therapy failures. Both are meant to commit the SAME assigned misstep. You do NOT know which is which — judge blind. Each item in the job gives: ‘client‘ (what the client said), ‘criteria‘ (the skill a GOOD response would show), ‘display_name‘ + ‘definition‘ (the assigned misstep), and ‘A‘ / ‘B‘ (the two candidate bad responses). Pick the **better exemplar for the dataset**, judging in this priority order: 1. **Plausibility** — could a real, well-meaning but unskillful therapist actually say this out loud in session, word for word? Stilted, robotic, or ”AI-sounding” prose is a defect. 2. **Category fidelity** — does it clearly commit THIS misstep as its main move (not a neighboring error, not generic badness)? 3. **Natural spoken voice** — warmth/fluency on the surface with the badness in the therapeutic MOVE is ideal; canned phrasing or a cold register is worse.

Generation Stage	Prompt Description	Prompt
		If they are genuinely equal, pick the one you'd rather have in the dataset. You must choose A or B — no ties. ## Output (STRICT) Write ONLY a raw JSON array to the output file, one object per item, in order: “‘json [{”sid”: ”ab0001”, ”winner”: ”A”, ”reason”: ”<=12 words”}, {”sid”: ”ab0002”, ”winner”: ”B”, ”reason”: ”...”}] “‘ No prose, no markdown fences. Every ‘sid‘ in the job must appear exactly once.
	#2.5. Real example judge request comparing previously written unskilled therapist response and new unskilled response	CLIENT: [concerned] I’d like to exercise more. I hate the way I look and feel. CRITERIA (skill a good response would show): The clinician guesses at a deeper meaning in the client's words. Reflections are offered as statements, not questions. The reflection does not include the clinician's opinion, advice, or information. MISSTEP: Advice-giving -- Giving advice or using direct persuasion to tell the recipient what to do without teaching, reinforcing, or consolidating a generalizable skill.
	#2.6. Real example judge response comparing previously written unskilled therapist response and new unskilled response	{”sid”: ”ab0086”, ”exercise_id”: ”mi__s2__example__e2”, ”v2_is”: ”A”, ”A”: ”Get yourself into a gym this week and exercise three mornings before work.”, ”B”: ”Sign up for a gym this week and commit to three mornings before work. That is how you turn this around.”} # ab_verdicts_claude.jsonl (verbatim) {”sid”: ”ab0086”, ”winner”: ”B”}

D Accounting for textual confounds

Table 1 shows linear separation through held-out AUC between skilled therapist responses from the deliberate practice books (Appendix A) and the unskilled therapist responses generated by an LLM (Appendix C). Two potential confounds could account for that linear separation: 18% median difference in textual confounds, and the human vs. LLM generated source of the text groups. To account for these differences, we generate skilled therapist responses using an LLM, and then compare them with the existing LLM-generated unskilled therapist response.

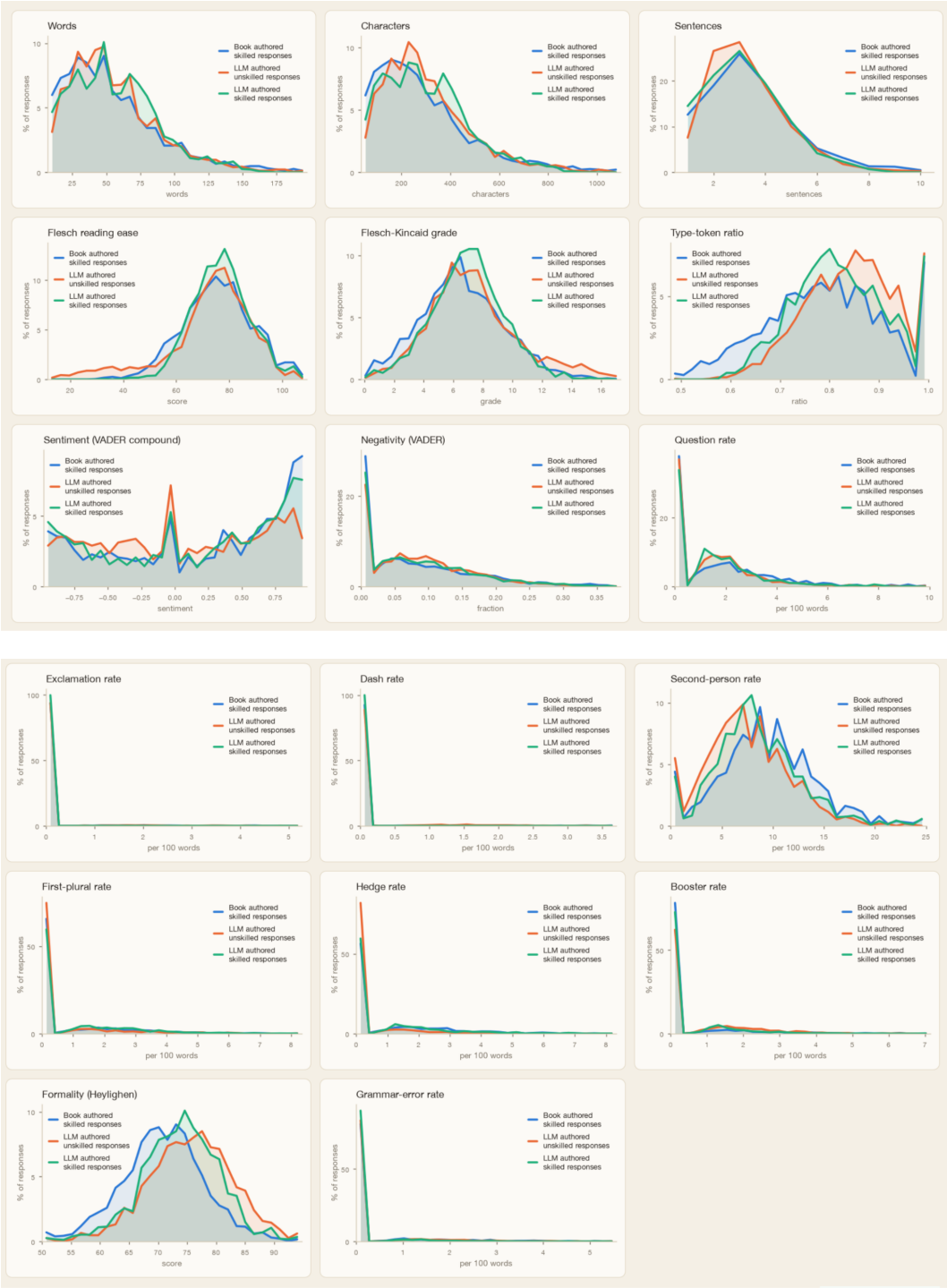


Figure 8. Rates of seventeen textual confounds across all three datasets. Rates are expressed as percentage of the dataset with specific values of textual confounds.

Table 11. List of textual confounds studied including: name, definition, an LLM-authored plain english description, sample code, and mean values. The mean values in "Book authored skilled responses" belong to the textual confounds in the skilled therapist responses from the books. The mean values in "LLM authored unskilled responses" are for the unskilled therapist responses authored in appendix C, and are compared in percentages against the previous column. The mean values in "LLM authored skilled responses" were authored specifically for this section, and compared against the previous column. It should be noted "LLM authored skilled responses" were not used in GoodBadTherapy dataset, just this section.

Metric	What it measures (plain English)	How it is counted	Book authored skilled responses	LLM authored unskilled responses	LLM authored skilled responses
Words	How many words the response has.	len(re.findall(r"[A-Za-z0-9']+", text))	58.15	58.18 (+0.06%)	56.29 (-3.25%)
Characters	How many characters (letters, spaces, punctuation).	len(text)	313.82	319.25 (+1.73%)	301.39 (-5.59%)
Sentences	How many sentences (split on . ! ?).	SENTENCE_SPLIT = re.compile(r"[.!?]+") len(SENTENCE_ SPLIT.split(text))	3.72	3.39 (-8.78%)	3.17 (-6.46%)
Flesch reading ease	Readability, 0-100, higher = easier to read.	206.835 - 1.015* words/sents - 84.6* syllables/words	75.85	71.26 (-6.05%)	77.09 (+8.18%)
Flesch–Kincaid grade	US school grade level needed to read it (higher = harder).	0.39*words/sents + 11.8*syllables/words - 15.59	6.55	7.51 (+14.68%)	6.92 (-7.81%)
Type–token ratio	Vocabulary variety in this one response: unique words / total words (higher = less repetition). Shrinks with length.	len(set(tokens)) / len(tokens)	0.78	0.84 (+7.35%)	0.82 (-2.33%)
Sentiment (VADER compound)	Overall sentiment, -1 (very negative) to +1 (very positive).	vader.polarity_ scores(text)["compound"]	0.19	0.06 (-70.29%)	0.15 (+164.82%)
Negativity (VADER)	Fraction of the text that reads negative in tone.	vader.polarity_ scores(text)["neg"]	0.09	0.09 (+0.10%)	0.09 (-2.25%)
Question rate	Question marks per 100 words (asking vs telling).	100*text.count("?")/ words	1.88	1.52 (-18.87%)	1.61 (+5.48%)
Exclamation rate	Exclamation marks per 100 words.	100*text.count("!")/ words	0.21	0.17 (-20.57%)	0.01 (-93.35%)
Dash rate	Em/en dashes per 100 words (an 'AI-style' punctuation tell). [Artificially set to 0 in LLM authored skilled responses using prompt guidance]	100*(text.count("-")+ text.count("-"))/ words	0.17	0.21 (+21.28%)	0.00 (-100.00%)

Metric	What it measures (plain English)	How it is counted	Book authored skilled responses	LLM authored unskilled responses	LLM authored skilled responses
Second-person rate	'you/your' words per 100 words (addressing the client directly).	SECOND_PERSON = re.compile(r"\b(you|your|yours|yourself)\b", re.I) 100*len(SECOND_PERSON.findall(text))/words	9.59	7.47 (-22.05%)	8.54 (+14.27%)
First-person plural rate	'we/us/our' words per 100 words (collaborative language).	FIRST_PLURAL = re.compile(r"\b(we|us|our|ours|let's|lets)\b", re.I) 100*len(FIRST_PLURAL.findall(text))/words	0.99	0.70 (-29.99%)	1.08 (+54.67%)
Hedge rate	Softening/tentative words (maybe, perhaps, might, it sounds like) per 100 words.	HEDGE_TERMS = ["maybe","perhaps","it seems","i wonder","sort of","kind of","might","could","possibly","i guess","i think","it sounds like"] 100*sum(text.lower().count(t) for t in HEDGE_TERMS)/words	1.22	0.37 (-69.99%)	1.02 (+179.44%)
Booster rate	Emphatic/certainty words (definitely, obviously, clearly, really) per 100 words.	BOOSTER_TERMS = ["definitely","obviously","clearly","honestly","frankly","absolutely","certainly","really","literally","simply"] 100*sum(text.lower().count(t) for t in BOOSTER_TERMS)/words	0.58	0.86 (+48.41%)	0.56 (-35.21%)
Formality	Heylighen formality, 0-100: high when the text leans on nouns/adjectives/prepositions (informational), low when it leans on pronouns/verbs/adverbs (conversational).	50*((noun+adj+prep+art - (pron+verb+adv+intj))/total + 1)	70.67	75.75 (+7.19%)	73.99 (-2.33%)
Grammar-error rate	Grammar issues per 100 words (LanguageTool).	100*len(language_tool.check(text))/words	0.45	0.36 (-18.10%)	0.18 (-51.48%)

Table 12. Means of seventeen textual confounds in each of the three datasets, a description, and code used to calculate them. The % difference in "LLM authored unskilled responses" column is between that column and the "Book authored skilled responses" column, which is the text confounding rate in the GoodBadTherapy dataset. The % difference in "LLM authored skilled responses" column is with the "LLM authored unskilled responses" column, representing mitigating textual confounds.

The GoodBadTherapy dataset (comparing book-authored skilled examples vs. LLM-authored unskilled examples) had an 18.1% median textual confounds across 17 categories. However, by changing the comparison to LLM-authored skilled examples vs. LLM-authored unskilled examples, the median percentage of textual confounds dropped from 18.1% to 8.2%.

Models	**GoodBadTherapy: book-authored skilled examples vs. LLM-authored unskilled examples**	**Removing confounds: LLM-authored skilled examples vs. LLM-authored unskilled examples**
Textual confounds Median % across all models	18.1%	8.2%

Table 13. The GoodBadTherapy dataset (comparing book-authored skilled examples vs. LLM-authored unskilled examples) had an 18.1% median textual confounds across 17 categories. However, by changing the comparison to LLM-authored skilled examples vs. LLM-authored unskilled examples, the median percentage of textual confounds dropped from 18.1% to 8.2%.

After controlling for book-authored vs. LLM-authored and reducing textual confounds median from 18.1% to 8.2%, the held-out AUC is still near-perfect, demonstrating the source of linear separation is unlikely to be from textual confounds or human vs LLM text source.

Model	**Linear probe Held-out AUC on GoodBadTherapy:** book-authored skilled examples vs. LLM-authored unskilled examples.	**Linear probe Held-out AUC after Removing confounds:** LLM-authored skilled examples vs. LLM-authored unskilled examples
Qwen3-4B	0.9997	**0.9977**
Qwen3-8B	0.9999	**0.9989**
Llama-3.1-8B	0.9999	**0.9989**
Gemma-3-12B	0.9999	**0.9989**
Gemma-3-27B	1.0000	**0.9995**
OLMo-3-7B	0.9998	**0.9982**

Table 14. Linear Probe held-out AUC is still near perfect even after reducing textual confounds and matching authoring source across both datasets.

E GoodBadTherapy dataset and Healing Axes

For replication purposes, the GoodBadTherapy dataset is provided as a public approval-required dataset on HuggingFace. Apologies to ICLR'27 reviewers, we recommend creating an anonymous account to access the repository to maintain double-blind review.

HuggingFace dataset link @ https://huggingface.co/datasets/ICLR-2027-13513/HealingAxis_GoodBadTherapy/tree/main

The Healing Axes safetensors files derived from GoodBadTherapy dataset contrasting pairs can be found on a public GitHub repo @ https://github.com/ICLR-2027-13513/HealingAxis

F SkilledTherapyBenchmark

For replication purposes, the entire prompt used in LLM-as-judge code is provided under the section3_SkilledTherapyBenchmark folder. The results of evaluating the GoodBadTherapy dataset using *SkilledTherapyBenchmark* are available under the appendixF_SkilledTherapyBenchmark_on_GoodBadTherapy folder.

These files can be found on a public GitHub repo @ https://github.com/ICLR-2027-13513/HealingAxis

Sample judge prompt

Table 15. *SkilledTherapyBenchmark* prompt sent to GLM-5.2 no-reasoning as the judge prompt for deliberate practice therapy skills. Inputs are the client utterance and therapist response. Output is a JSON outlining which specific skill criteria were met, as a result which skills were met, and then finally the number of skills met.

»»> SYSTEM MESSAGE (stable cached prefix: role + output contract + all 130 skills / 378 criteria) »»>
You are a clinical judge grading one therapist response against the numbered criteria of many psychotherapy skills at once. You are given a single client and therapist exchange, then a checklist that lists every skill with its numbered criteria. Judge that one exchange against every criterion of every skill below.
Judge each criterion independently. A doubt about one criterion must not lower your verdict on another. A single short exchange can only exercise a few skills, so most skills will have no occasion here. That is expected and is handled by the verdicts below.
VERDICTS (assign exactly one per criterion):
- TRUE: the response performs this criterion's requirement, even briefly, implicitly, or in different words. Credit a genuine attempt; do not require the ideal or fullest form.
- FALSE: the criterion applies here and the response fails it: absent, merely generic, the opposite, or a required element omitted.
- N/A: assign strictly per the N/A POLICY below.
N/A POLICY: Use N/A only when a criterion is (a) genuinely unobservable from a text transcript, for example eye contact, forward lean, or tone of voice, or (b) there is genuinely no occasion to exercise it in this single exchange. Never use N/A (or TRUE) to excuse a response that plainly does the opposite of the criterion. That is what FALSE is for. When a skill is genuinely in play its criteria cannot all be N/A, but a skill that has no occasion at all in this exchange may be entirely N/A.
OUTPUT FORMAT: return ONLY a single JSON object, with no prose, no explanation, and no markdown code fences. Use exactly this shape:
{”skills”:[{”id”:”<skill_id>”,”v”:”<letters>”,”met”:<0 or 1>}, ...]}
Rules for the JSON:
- Include exactly one entry for every skill in the SKILL CRITERIA CHECKLIST below, in the same order they appear.
- ”id” is the skill id exactly as printed in that skill's header, for example ”mi:1”.
- ”v” is a string of single letter verdicts, one letter per criterion in the listed order: T for TRUE, F for FALSE, N for N/A. The number of letters must equal that skill's criterion count.
- ”met” is 1 if the response demonstrates the skill, otherwise 0. Compute it by this rule: met is 1 when the skill's verdict letters contain zero F and at least one T; met is 0 otherwise, meaning any F is present or every letter is N.
========== SKILL CRITERIA CHECKLIST (130 skills, 378 criteria) ==========
Judge the single exchange in the user message against every criterion listed below.
aedp:1 - Moment-to-Moment Tracking (Accelerated Experiential Dynamic Psychotherapy (AEDP))
0. Name in a statement a nonverbal/physical movement of your client. Make your statement descriptive and short.
<... 128 skills …>
systemic:12 - Interactional Structuring: Initiating Enactments (Systemic Family Therapy)
0. Reflect or validate the underlying meaning or emotion of what Client 1 just said.
1. Let Client 1 know that you want them to turn and speak directly to Client 2.
2. Select one of the following ways to initiate the enactment:
»»> USER MESSAGE (per-exchange; the ONLY part that changes) »»>
Client: {CLIENT}
Therapist: {THERAPIST}
Grade the exchange above against every skill's criteria. Return only the JSON.

Sample response

Table 16. JSON output for the *SkilledTherapyBenchmark* judge. Each skill returns the skill identifier, the verdict as True, False or N/A, and whether the skill was met.


```
{"skills":[{"id":"aedp:1","v":"N","met":0},{"id":"aedp:2","v":"N","met":0},{"id":"aedp:3","v":
"N","met":0},{"id":"aedp:4","v":"N","met":0},{"id":"aedp:5","v":"N","met":0},{"id":"aedp:6",
"v":"N","met":0},{"id":"aedp:7","v":"N","met":0},{"id":"aedp:8","v":"N","met":0},{"id":"aedp:
9","v":"N","met":0},{"id":"aedp:10","v":"N","met":0},{"id":"aedp:11","v":"N","met":0},{"id":
"aedp:12","v":"N","met":0},{"id":"cbt:1","v":"N","met":0},{"id":"cbt:2","v":"N","met":0},
{"id":"cbt:3","v":"N","met":0},{"id":"cbt:4","v":"N","met":0},{"id":"cbt:5","v":"N","met":0},
{"id":"cbt:6","v":"N","met":0},{"id":"cbt:7","v":"N","met":0},{"id":"cbt:8","v":"N","met":0},
{"id":"cbt:9","v":"N","met":0},{"id":"cbt:10","v":"N","met":0},{"id":"efct:1","v":"N","met":
0},{"id":"efct:2","v":"N","met":0},{"id":"efct:3","v":"N","met":0},{"id":"efct:4","v":"N",
"met":0},{"id":"efct:5","v":"N","met":0},{"id":"efct:11","v":"N","met":0},{"id":"efct:12","v":
"N","met":0},{"id":"eft:1","v":"N","met":0},{"id":"eft:2","v":"N","met":0},{"id":"eft:3","v":
"N","met":0},{"id":"eft:4","v":"N","met":0},{"id":"eft:5","v":"N","met":0},{"id":"eft:6","v":
"N","met":0},{"id":"eft:7","v":"N","met":0},{"id":"eft:8","v":"N","met":0},{"id":"eft:9","v":
"N","met":0},{"id":"eft:10","v":"N","met":0},{"id":"eft:11","v":"N","met":0},{"id":"eft:12",
"v":"N","met":0},{"id":"ipt:1","v":"N","met":0},{"id":"ipt:2","v":"N","met":0},{"id":"ipt:3",
"v":"N","met":0},{"id":"ipt:4","v":"N","met":0},{"id":"ipt:5","v":"N","met":0},{"id":"ipt:6",
"v":"N","met":0},{"id":"ipt:8","v":"N","met":0},{"id":"ipt:9","v":"N","met":0},{"id":"ipt:10",
"v":"N","met":0},{"id":"mi:1","v":"TTTT","met":1},{"id":"mi:2","v":"FTF","met":0},{"id":"mi:
3","v":"FTF","met":0},{"id":"mi:4","v":"FF","met":0},{"id":"mi:5","v":"N","met":0},{"id":"mi:
6","v":"N","met":0},{"id":"mi:7","v":"N","met":0},{"id":"mi:8","v":"N","met":0},{"id":"mi:9",
"v":"N","met":0},{"id":"mi:10","v":"N","met":0},{"id":"multicultural:1","v":"N","met":0},
{"id":"multicultural:2","v":"N","met":0},{"id":"multicultural:3","v":"N","met":0},{"id":
"multicultural:4","v":"N","met":0},{"id":"multicultural:5","v":"N","met":0},{"id":
"multicultural:6","v":"N","met":0},{"id":"multicultural:7","v":"N","met":0},{"id":
"multicultural:8","v":"N","met":0},{"id":"multicultural:9","v":"N","met":0},{"id":
"multicultural:10","v":"N","met":0},{"id":"multicultural:11","v":"N","met":0},{"id":
"multicultural:12","v":"N","met":0},{"id":"psychedelic:1","v":"N","met":0},{"id":"psychedelic:
2","v":"N","met":0},{"id":"psychedelic:3","v":"N","met":0},{"id":"psychedelic:4","v":"N",
"met":0},{"id":"psychedelic:5","v":"N","met":0},{"id":"psychedelic:6","v":"N","met":0},{"id":
"psychedelic:7","v":"N","met":0},{"id":"psychedelic:8","v":"N","met":0},{"id":"psychedelic:9",
"v":"N","met":0},{"id":"psychedelic:10","v":"N","met":0},{"id":"psychedelic:11","v":"N","met":
0},{"id":"psychedelic:12","v":"N","met":0},{"id":"psychodynamic:1","v":"N","met":0},{"id":
"psychodynamic:3","v":"N","met":0},{"id":"psychodynamic:4","v":"N","met":0},{"id":
"psychodynamic:5","v":"N","met":0},{"id":"psychodynamic:6","v":"N","met":0},{"id":
"psychodynamic:7","v":"N","met":0},{"id":"psychodynamic:8","v":"N","met":0},{"id":
"psychodynamic:9","v":"N","met":0},{"id":"psychodynamic:11","v":"N","met":0},{"id":
"psychodynamic:12","v":"N","met":0},{"id":"rebt:1","v":"N","met":0},{"id":"rebt:2","v":"N",
"met":0},{"id":"rebt:3","v":"N","met":0},{"id":"rebt:4","v":"N","met":0},{"id":"rebt:5","v":
"N","met":0},{"id":"rebt:6","v":"N","met":0},{"id":"rebt:7","v":"N","met":0},{"id":"rebt:8",
"v":"N","met":0},{"id":"rebt:9","v":"N","met":0},{"id":"rebt:10","v":"N","met":0},{"id":"rebt:
11","v":"N","met":0},{"id":"rebt:12","v":"N","met":0},{"id":"schema:1","v":"N","met":0},{"id":
"schema:2","v":"N","met":0},{"id":"schema:3","v":"N","met":0},{"id":"schema:4","v":"N","met":
0},{"id":"schema:5","v":"N","met":0},{"id":"schema:6","v":"N","met":0},{"id":"schema:7","v":
"N","met":0},{"id":"schema:8","v":"N","met":0},{"id":"schema:9","v":"N","met":0},{"id":
"schema:10","v":"N","met":0},{"id":"schema:11","v":"N","met":0},{"id":"schema:12","v":"N",
"met":0},{"id":"systemic:1","v":"N","met":0},{"id":"systemic:2","v":"N","met":0},{"id":
"systemic:3","v":"N","met":0},{"id":"systemic:4","v":"N","met":0},{"id":"systemic:5","v":"N",
"met":0},{"id":"systemic:6","v":"N","met":0},{"id":"systemic:7","v":"N","met":0},{"id":
"systemic:8","v":"N","met":0},{"id":"systemic:9","v":"N","met":0},{"id":"systemic:10","v":"N",
"met":0},{"id":"systemic:11","v":"N","met":0},{"id":"systemic:12","v":"N","met":0}]}
```


Confusion Matrix on GoodBadTherapy dataset

Table 17. Confusion matrix for *SkilledTherapyBenchmark* LLM judge. GOOD here is a skilled therapist response, and BAD here is an unskilled therapist response. TP/True-Positives are skilled therapist responses from the books that are correctly identified as skilled by the judge. FN/False-Negatives are skilled therapist responses from the books that are incorrectly identified as unskilled by the judge. FP/False-Positives are unskilled therapist responses authored by an LLM with a therapist misstep that were incorrectly identified by the judge as skilled. TN/True-Negatives are unskilled therapist responses authored by an LLM with a therapist misstep that were correctly identified as not meeting the skill criteria.

	Predicted GOOD (met)	**Predicted BAD (not met)**
Actual GOOD	**TP 1675**	**FN 634**
Actual BAD	**FP 167**	**TN 2128**

Additional Judge Metrics

Table 18. *SkilledTherapyBenchmark* classification metrics on the *GoodBadTherapy* dataset: F1, accuracy, precision, sensitivity, specificity, sample size, and total number of majority@3 LLM calls.

F1	**80.7%**
Accuracy	82.6%
Precision	90.9% of responses judged skilled are skilled
Sensitivity / recall	72.5% of skilled responses are judged skilled
Specificity	92.7% of unskilled responses judged as unskilled
N	4,604 (2,309 skilled responses, 2,295 unskilled responses)
majority@3 Total Judge calls	13,812

G Alpha and Layer Sweep

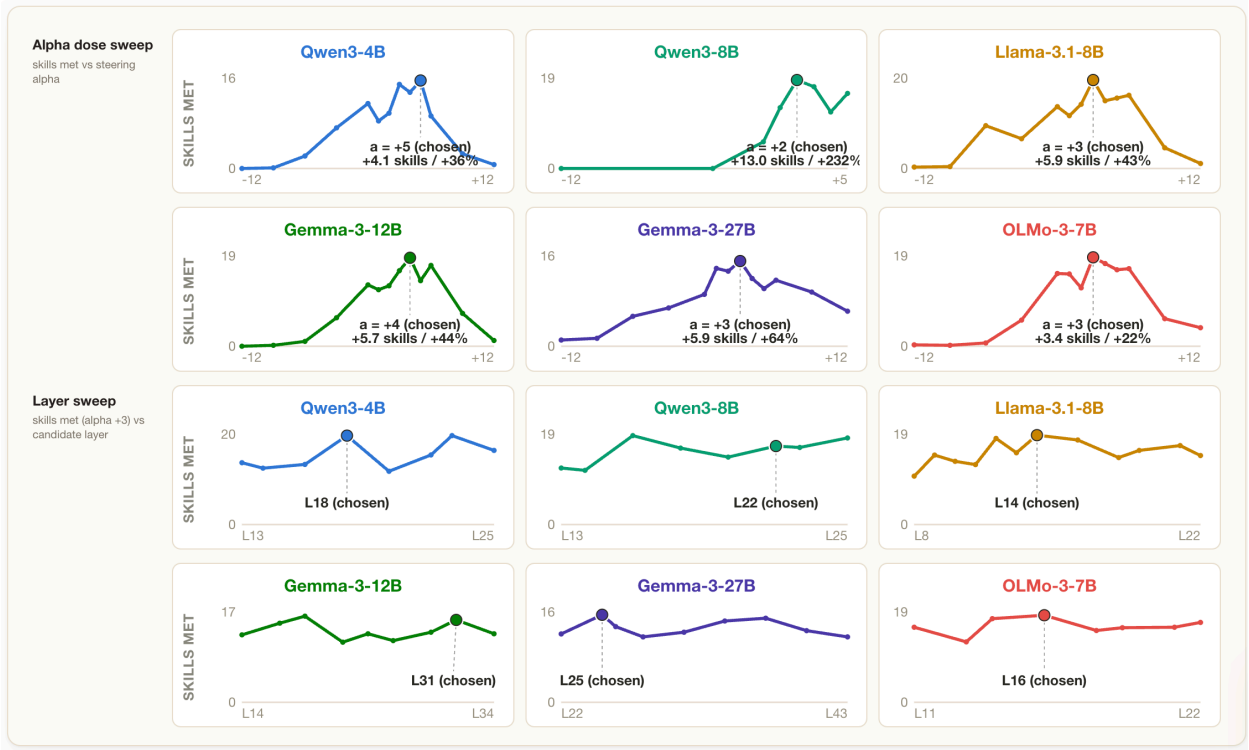


Figure 9. Activation-steering sweeps for all six models. **Top two rows:** alpha sweep. Number of skills met (Y-axis, out of 130 skills) against specific alpha values used in the sweep (X-axis). **Bottom two rows**: layer sweep. Number of skills met (Y-axis, out of 130 skills) at alpha=3 against specific layers evaluated. The final chosen/ideal layer and alpha for each model are marked with a filled circle. This figure is the uncropped version of Figure 3 in Section 3.2.

When using activation steering it's critical to identify the ideal activation steering alpha (strength) in the ideal layer. In order to find those ideal values for each model, we first use the *SkilledTherapyBenchmark* in a layer sweep using alpha=3 to find the layer with the maximum clinical skills. Once the ideal layer has been identified, we perform an alpha sweep in that layer. Table 19 lists the results of the layer sweep, and Table 20 lists the results of the subsequent alpha sweep in those layers. Table 21 summarizes these findings.

Table 19. Layer sweep results checking for the number of clinical skills demonstrated in each layer for each model with activation steering at alpha=3. The number of clinical skills is determined by using the *SkilledTherapyBenchmark* (Appendix F). The baseline (0% improvement) is the argmax layer for the linear probe.

Layer	Qwen3-4B	Qwen3-8B	Llama-3.1-8B	Gemma-3-12B	Gemma-3-27B	OLMo-3-7B
L8			10.2 (-44%)			
L9			14.6 (-19%)			
L10			13.3 (-27%)			
L11			12.6 (-31%)			15.8 (-10%)
L12			18.1 (0%)			
L13	13.7 (-11%)	11.9 (-28%)	15.1 (-17%)			12.7 (-28%)
L14	12.5 (-19%)	11.4 (-31%)	**18.8 (+4%)**	12.7 (0%)		17.6 (0%)
L16	13.3 (-14%)	18.7 (+13%)	17.8 (-2%)			**18.3 (+4%)**

Layer	Qwen3-4B	Qwen3-8B	Llama-3.1-8B	Gemma-3-12B	Gemma-3-27B	OLMo-3-7B
L17				14.9 (+17%)		
L18	**19.7 (+28%)**	16.1 (-3%)	14.1 (-22%)			15.1 (-14%)
L19			15.6 (-14%)	16.2 (+28%)		15.7 (-11%)
L20	11.8 (-24%)	14.2 (-14%)				
L21			16.6 (-8%)			15.8 (-10%)
L22	15.4 (0%)	**16.5 (0%)**	14.5 (-20%)	11.3 (-11%)	12.1 (-9%)	16.8 (-5%)
L23	19.7 (+28%)	16.2 (-2%)				
L24				12.9 (+1%)		
L25	16.4 (+7%)	18.2 (+10%)			**15.5 (+16%)**	
L26				11.6 (-9%)	13.4 (0%)	
L28					11.6 (-13%)	
L29				13.2 (+4%)		
L31				**15.5 (+22%)**	12.4 (-7%)	
L34				12.9 (+2%)	14.4 (+8%)	
L37					14.9 (+11%)	
L40					12.7 (-5%)	
L43					11.6 (-13%)	

Table 20. Alpha sweep results checking for the number of clinical skills demonstrated at each alpha for each model with activation steering. The number of clinical skills is determined by using the *SkilledTherapyBench-mark* (Appendix F). The baseline (0% improvement) is the alpha=0 value indicating no activation steering.

Alpha	Qwen3-4B	Qwen3-8B	Llama-3.1-8B	Gemma-3-12B	Gemma-3-27B	OLMo-3-7B
-12	0.0 (-100%)	0.0 (-100%)	0.3 (-98%)	0.0 (-100%)	1.1 (-88%)	0.3 (-98%)
-9	0.1 (-99%)		0.4 (-97%)	0.2 (-98%)	1.4 (-85%)	0.2 (-99%)
-6	2.2 (-81%)		9.5 (-31%)	1.0 (-92%)	5.3 (-42%)	0.7 (-96%)
-3	7.2 (-38%)	0.0 (-100%)	6.6 (-52%)	6.0 (-54%)	6.8 (-26%)	5.5 (-64%)
0	11.5 (0%)	5.6 (0%)	13.7 (0%)	12.9 (0%)	9.2 (0%)	15.3 (0%)
+1	8.4 (-27%)	12.8 (+128%)	11.7 (-15%)	11.9 (-7%)	13.8 (+51%)	15.2 (0%)
+2	9.8 (-15%)	**18.6 (+230%)**	14.2 (+3%)	12.7 (-2%)	13.3 (+45%)	12.3 (-19%)
+3	14.9 (+29%)	17.2 (+205%)	**19.6 (+43%)**	15.9 (+23%)	**15.1 (+65%)**	**18.7 (+22%)**
+4	13.5 (+17%)	11.9 (+111%)	15.0 (+9%)	**18.6 (+45%)**	12.0 (+31%)	17.4 (+14%)
+5	**15.6 (+35%)**	15.8 (+179%)	15.6 (+14%)	13.8 (+7%)	10.2 (+11%)	16.1 (+5%)
+6	9.3 (-20%)		16.2 (+18%)	17.0 (+32%)	11.7 (+28%)	16.3 (+7%)
+9	2.6 (-77%)		4.6 (-67%)	6.9 (-47%)	9.6 (+5%)	5.8 (-62%)
+12	0.7 (-94%)		1.1 (-92%)	1.2 (-90%)	6.2 (-32%)	3.9 (-75%)

Table 21. Summary of activation steering sweeps across models locating ideal alpha and layer for each model.

Model	Ideal alpha	Ideal layer	Skills-met delta at ideal alpha and layer
Qwen3-4B	+5	L18	+4.1
Qwen3-8B	+2	L22	+13.0
Llama-3.1-8B	+3	L14	+5.9
Gemma-3-12B	+4	L31	+5.8
Gemma-3-27B	+3	L25	+6.0
OLMo-3-7B	+3	L16	+3.4

H Withheld Modality Healing Axes

This is the appendix for Section 3.3's results on re-deriving the Healing Axis by leaving one book out, and how well that modified axis performs on the withheld modality. The withheld modality re-derived axes were used in activation steering over six open models on client utterances from their respective withheld books. The tables below show the difference in the number of clinical skills observed by the *SkilledTherapyBenchmark* judge, between activation steering along modality-withheld Healing Axes and unsteered baseline.

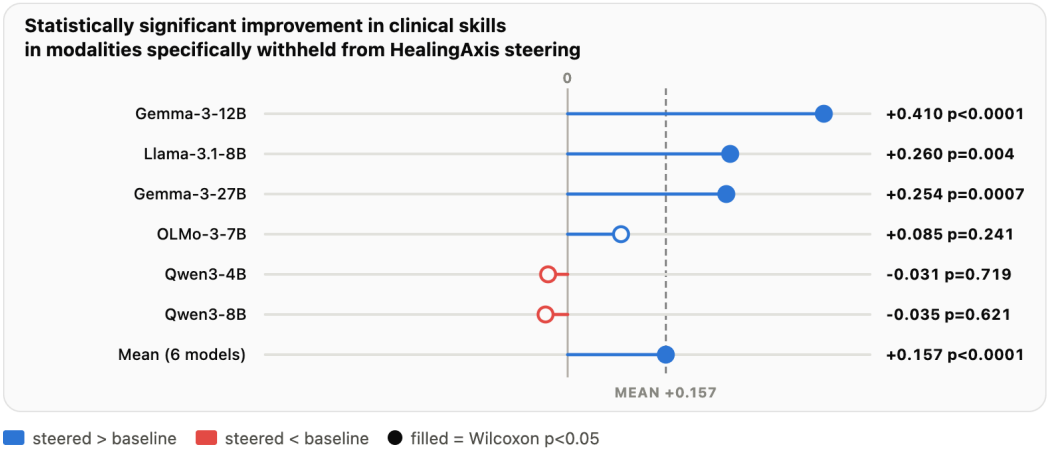


Figure 10. The difference in the number of clinical skills observed by *SkilledTherapyBenchmark* between unsteered baseline and activation steering along modality-withheld Healing Axes. Evaluated on client utterances from the withheld books. Filled markers are statistically significant with p<0.05, and unfilled markers have a p>0.05. Four models have directionally positive results, of which three are statistically significant. Both Qwen models were negative and not statistically significant.

Table 22. Results breakdown of comparing number of skills met in baseline/unsteered responses vs. activation steering along modality-withheld Healing Axis. *GoodBadTherapy* evaluations on withheld modality Healing Axis grouped by open model. Each of the dozen withheld-modality Healing Axis ran on N=40 consistent client utterances from that same withheld modality book for a total of 480 evaluations per model. Base count is the number of clinical skills met prior to steering, and Steered count is the number of skills with modality-withheld Healing Axis activation steering. "Δ count" is the difference between those.

model	n	Base count	Steered count	Δ count	win/loss/tie	Wilcoxon p	any-rate	McNemar p
Gemma-3-12B L31 α+4	480	0.902	1.312	**+0.410**	154/48/278	<1e-4	36.9%→51.7%	<1e-4
Llama-3.1-8B L14 α+3	480	1.227	1.488	**+0.260**	126/99/255	0.004	48.3%→52.3%	0.096

model	n	Base count	Steered count	Δ count	win/loss/tie	Wilcoxon p	any-rate	McNemar p	
Gemma-3-27B L25 α+3	480	1.883	2.138	**+0.254**	155/93/232	0.0007	58.1%→65.2%	0.0001	
OLMo-3-7B L16 α+3	480	1.167	1.252	**+0.085**	100/85/295	0.241	45.8%→51.5%	0.002	n.s.
Qwen3-4B L18 α+5	480	1.608	1.577	**-0.031**	113/136/231	0.719	58.8%→56.5%	0.260	n.s.
Qwen3-8B L22 α+2	480	1.354	1.319	**-0.035**	112/116/252	0.621	53.1%→50.6%	0.246	n.s.
Mean (all 6 models)	**2,880**	**1.357**	**1.514**	**+0.157**	**760/577/1,543**	**<1e-4**	**50.2%→54.6%**	**<1e-4**	

Per-modality breakdown

Table 23. Results breakdown of comparing number of skills met in baseline/unsteered responses vs. activation steering along modality-withheld Healing Axis in the same modality. Pooled across all six models and sorted descending by effect size.

modality	n	base	steer	Δ count	win/loss/tie	Wilcoxon p	McNemar p
Multicultural / culturally-responsive	240	2.400	2.842	**+0.442**	106/63/71	0.0003 *	<1e-4
IPT Interpersonal Psychotherapy	240	1.038	1.446	**+0.408**	104/40/96	<1e-4 *	0.0001
Systemic / family therapy	240	1.996	2.383	**+0.388**	95/72/73	0.092	0.041
CBT Cognitive Behavioral Therapy	240	3.671	3.942	**+0.271**	85/70/85	0.137	0.099
MI Motivational Interviewing	240	0.704	0.933	**+0.229**	56/35/149	0.003 *	0.519
Psychodynamic therapy	240	0.271	0.338	**+0.067**	41/27/172	0.158	0.162
Psychedelic-assisted therapy	240	1.904	1.954	**+0.050**	86/78/76	0.510	0.072
Schema Therapy	240	1.825	1.875	**+0.050**	77/77/86	0.616	1.000
EFT Emotion-Focused Therapy	240	0.279	0.325	**+0.046**	29/23/188	0.365	0.644
EFCT Emotionally Focused Couple Therapy	240	0.096	0.117	**+0.021**	18/13/209	0.502	0.585
REBT Rational Emotive Behavior Therapy	240	0.017	0.013	**-0.004**	2/3/235	1.000	1.000
AEDP Accelerated Experiential Dynamic Psychotherapy	240	2.083	2.004	**-0.079**	61/76/103	0.303	0.868

Settings used in leave-one-modality-out steering

Table 24. Specific alpha/strength and layer used in activation steering for the modality-withheld *SkilledTherapyBenchmark* evaluations. Includes number of evaluations performed per model.

model	steer layer	dose α	n
Gemma-3-12B	**L31**	**α+4**	480
Llama-3.1-8B	**L14**	**α+3**	480
Gemma-3-27B	**L25**	**α+3**	480
OLMo-3-7B	**L16**	**α+3**	480
Qwen3-4B	**L18**	**α+5**	480
Qwen3-8B	**L22**	**α+2**	480

I KNN Clustering

This appendix contains the specific results behind k-means / KNN clustering sweeps used in Section 4.1. The sweep evaluates the silhouette scores of specific clusters of skills in each model. Higher silhouette scores indicate more cohesive clusters. The k-means sweep ran on k values between 2 and 12.

An N-sweep between 2 and 12 for KNN clustering skills within the same book finds a global mean peak at N=3. Models exhibit a dual peak structure where the first peak is meta-therapeutic techniques, and a second primary peak effectively mirrors per-book reconstruction.

Table 25. Mean silhouette cosine score for each model (rows) for each potential cluster count k from 2 to 12 (columns). The bottom row is the mean pooled over all models. In bold is the highest model-pooled mean value which occurs at k=3.

model	N2	N3	N4	N5	N6	N7	N8	N9	N10	N11	N12
Qwen3-4B	0.146	0.149	0.128	0.116	0.118	0.104	0.135	0.161	0.123	0.086	0.101
Qwen3-8B	0.149	0.145	0.120	0.136	0.141	0.087	0.150	0.121	0.122	0.106	0.112
Llama-3.1-8B	0.124	0.138	0.139	0.109	0.146	0.141	0.138	0.111	0.132	0.136	0.123
Gemma-3-12B	0.114	0.120	0.120	0.134	0.117	0.133	0.105	0.106	0.097	0.101	0.124
Gemma-3-27B	0.114	0.127	0.135	0.148	0.144	0.141	0.096	0.131	0.085	0.146	0.121
OLMo-3-7B	0.131	0.133	0.113	0.119	0.110	0.083	0.156	0.114	0.103	0.138	0.090
mean	0.130	**0.135**	0.126	0.127	0.129	0.115	0.130	0.124	0.110	0.119	0.112

J Sub-axis skills clustering

Using the k=3 number of clusters decided in Appendix I, which axis is each skill mapped to in each model? For all 130 clinical skills, we list below which axis is assigned in each model. Sub-axes are re-identified across models by matching the largest skills overlap. Each skill is counted as *unanimous* if all six open models assign it to the same sub-axis, and *majority* if four or more of the six open models assign it to the same sub-axis.

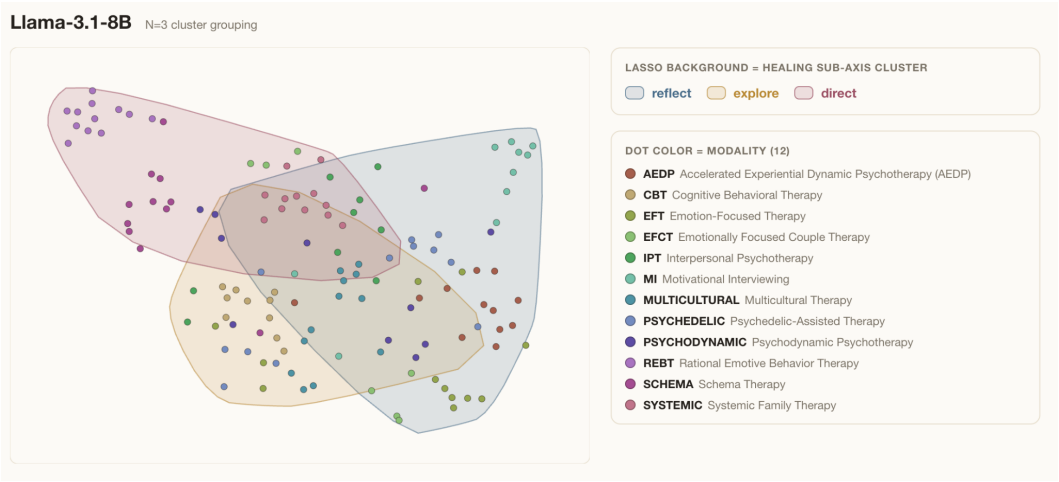


Figure 11. 2D t-SNE chart of skills within Llama-3.1-8B.

Table 26. Table listing for each of the 130 skills which sub-axis each model assigned it to. And a "consensus" row on how many models voted for the most common axis for that skill. Contains a summary row per book on how many skills were unanimous or majority votes across models. R stands for Reflect sub-axis; E stands for Explore sub-axis; D stands for Direct sub-axis. For example the top left cell signifies that AEDP's "Affirmative Work With Defenses" was assigned the Reflect sub-axis by Qwen3-4B.

skill	Qwen3-4B	Qwen3-8B	Llama-3.1-8B	Gemma-3-12B	Gemma-3-27B	OLMo-3-7B	consensus
Accelerated Experiential Dynamic Psychotherapy (AEDP) | 12 skills | 2 unanimous, 10 majority							
Affirmative Work With Defenses	R	D	R	E	E	D	R2/6
Affirming	E	R	R	R	R	R	R5/6
Exploring and Staying With Physical Experience	E	E	R	E	R	E	E4/6
Initiating Portrayals	E	E	R	E	R	R	R3/6
Metaprocessing	E	R	R	E	R	R	R4/6
Metatherapeutic Processing	E	E	R	E	R	E	E4/6
Moment-to-Moment Tracking	E	R	R	R	R	R	R5/6
Privileging Transformance Strivings	E	E	R	E	E	R	E4/6
Self-Involving Self-Disclosure	E	R	R	R	R	R	R5/6
Self-Revealing Self-Disclosure	E	R	R	R	R	R	R5/6
Undoing Aloneness	E	E	E	E	E	E	E6/6
Working With Anxiety	E	E	E	E	E	E	E6/6
Cognitive Behavioral Therapy (CBT) | 10 skills | 0 unanimous, 10 majority							
Adherence Flexibility	R	E	E	E	E	E	E5/6
Assigning and Reviewing Between-Session Activities	R	E	E	E	E	E	E5/6
Establishing Goals	R	E	E	E	E	E	E5/6
Explaining the Treatment Rationale for Cognitive Behavioral Therapy	R	E	E	E	E	E	E5/6
Negotiating a Session Agenda	R	E	E	E	E	E	E5/6
Responding to Client Resistance	R	E	E	E	E	E	E5/6
Responding to Therapeutic Alliance Ruptures	R	E	E	E	E	E	E5/6
Working With Behaviors	R	E	E	E	E	E	E5/6
Working With Cognitions	R	E	E	E	E	E	E5/6
Working With Emotions	R	E	E	E	E	E	E5/6
Emotion-Focused Therapy (EFT) | 12 skills | 1 unanimous, 11 majority							
Addressing Ruptures and Facilitating Repair	R	E	E	E	E	E	E5/6
Empathic Affirmation and Validation	E	R	R	R	R	R	R5/6
Empathic Conjectures	E	R	R	R	R	R	R5/6
Empathic Evocations	E	R	R	R	R	R	R5/6
Empathic Explorations	E	R	R	R	R	R	R5/6
Empathic Understanding	E	R	R	R	R	R	R5/6
Exploratory Questions	E	E	R	E	R	R	R3/6
Marker Recognition and Chair Work Task Setup	R	D	D	D	E	D	D4/6

skill	Qwen3-4B	Qwen3-8B	Llama-3.1-8B	Gemma-3-12B	Gemma-3-27B	OLMo-3-7B	consensus
Providing Treatment Rationale for Emotion-Focused Therapy	R	E	E	E	E	E	E5/6
Self-Disclosure	E	E	E	E	E	E	E6/6
Staying in Contact in the Face of Intense Affect	E	R	R	R	R	R	R5/6
Therapist Self-Awareness	E	R	E	R	R	R	R4/6
Emotionally Focused Couple Therapy (EFCT) | 7 skills | 3 unanimous, 6 majority							
Attachment-Reframed Validation	R	R	R	R	R	R	R6/6
Deepening Emotions With RISSSC: Speaking Simply, Slowly, and Softly	R	D	R	D	R	R	R4/6
Evocative Inquiry: Identifying Patterns and Eliciting Attachment Fears and Needs	R	D	D	D	E	D	D4/6
Evocative Reflection: Eliciting and Heightening Emotional Experience	R	R	R	R	R	R	R6/6
Interrupting Negative Process Early in Therapy: Catching the Bullet I	R	D	E	D	E	D	D3/6
Interrupting Negative Process Later in Therapy: Catching the Bullet II	R	D	D	D	E	D	D4/6
Validating Partners' Experiences and Tracking Dysfunctional Patterns	R	R	R	R	R	R	R6/6
Interpersonal Psychotherapy (IPT) | 9 skills | 1 unanimous, 6 majority							
Clarifications	R	E	R	R	R	R	R5/6
Exploring Interpersonal Relationships: Interpersonal Inventory II	R	E	E	E	E	E	E5/6
Generating Communication Options	R	E	R	E	E	D	E3/6
Helping the Client to Describe Their Distress and Need for Support	R	E	R	E	E	E	E4/6
Interpersonal Framing of Distress	R	R	R	R	R	R	R6/6
Mobilizing Social Support	R	E	R	E	E	E	E4/6
Motivating Interpersonal Action	R	E	R	E	E	R	R3/6
Presenting the Interpersonal Inventory: Interpersonal Inventory I	R	D	E	D	E	D	D3/6
Reinforcement of Effective Communication	R	E	R	E	E	E	E4/6
Motivational Interviewing (MI) | 10 skills | 7 unanimous, 8 majority							
Autonomy Support	R	E	R	E	E	E	E4/6
Complex Reflections, Part 1: Guesses at What the Client Means	R	R	R	R	R	R	R6/6
Complex Reflections, Part 2: Guesses at Underlying Client Emotions or Values	R	R	R	R	R	R	R6/6
Dancing With Discord	R	R	R	R	R	R	R6/6

skill	Qwen3-4B	Qwen3-8B	Llama-3.1-8B	Gemma-3-12B	Gemma-3-27B	OLMo-3-7B	consensus
Double-Sided Reflections	R	R	R	R	R	R	R6/6
Eliciting Change Talk	R	E	R	E	R	E	R3/6
Reflecting Change Talk	R	R	R	R	R	R	R6/6
Reflections and Open-Ended Questions	R	E	R	E	R	E	R3/6
Simple and Complex Affirmations	R	R	R	R	R	R	R6/6
Simple Reflections	R	R	R	R	R	R	R6/6
Multicultural Therapy (Multicultural) | 12 skills | 1 unanimous, 11 majority							
Acknowledging Therapist Limitations	R	E	E	E	E	E	E5/6
Assessing Client Expectations	R	E	E	E	E	E	E5/6
Gathering Information About Safety Concerns	R	E	R	E	E	E	E4/6
Inquiring About Cultural Implications of the Problem: Cultural Opportunities II	R	E	R	E	R	R	R4/6
Inquiring About Identity: Cultural Opportunities I	R	E	R	E	E	E	E4/6
Maintaining a Not-Knowing Stance: Cultural Humility II	R	E	E	E	E	E	E5/6
Reflecting Content Through a Cultural Lens	R	R	R	R	R	R	R6/6
Repairing Ruptures Due to Microaggressions	R	E	E	E	E	E	E5/6
Responding to Resistance and Ambivalence	R	E	E	E	E	E	E5/6
Talking About Sex and Success	R	E	R	E	E	R	R3/6
Therapist Self-Awareness: Cultural Humility I	E	D	E	E	E	D	E4/6
Working With Emotions in Context	R	E	R	E	E	E	E4/6
Psychedelic-Assisted Therapy (Psychedelic) | 12 skills | 0 unanimous, 9 majority							
Addressing Transference	R	E	E	E	E	E	E5/6
Boundaries and Informed Consent	R	E	E	E	E	E	E5/6
Compassionately Witnessing Strong Emotions	E	E	R	E	R	R	R3/6
Cultural Considerations: Racially and Ethnically Diverse Communities	R	E	E	E	E	E	E5/6
Embodying Insights: Integration III	R	E	R	E	E	E	E4/6
Exploring Intentions and Expectations	R	E	R	E	E	E	E4/6
Making Sense of the Experience: Integration I	R	E	R	E	E	E	E4/6
Navigating Strong Emotions	E	E	R	E	E	E	E5/6
Redirecting to the Body	E	E	R	E	R	R	R3/6
Responding to Relational Ruptures	R	E	E	E	E	E	E5/6
Working With Disappointment: Integration II	R	E	R	E	E	E	E4/6

skill	Qwen3-4B	Qwen3-8B	Llama-3.1-8B	Gemma-3-12B	Gemma-3-27B	OLMo-3-7B	consensus
Working With the Client's Internal Conflict	R	D	R	E	D	D	D3/6
Psychodynamic Psychotherapy (Psychodynamic) | 10 skills | 0 unanimous, 3 majority							
Deepening Emotional Experience	E	R	R	R	R	R	R5/6
Engaging in a Therapeutic Inquiry	R	E	R	E	R	E	R3/6
Exploring Fantasy	R	D	R	D	R	D	R3/6
Introducing the Rationale for Treatment	R	E	E	E	E	E	E5/6
Making Process Comments	R	D	E	E	E	D	E3/6
Making Transference Interpretations	R	D	R	E	E	D	R2/6
Pointing Out Defenses and Inquiring About Underlying Fear	R	E	R	E	E	E	E4/6
Providing a Corrective Emotional Experience	R	D	R	D	E	D	D3/6
Using Metaphors	E	E	R	E	R	R	R3/6
Using Supervision to Recognize Reenactments	R	D	E	D	E	D	D3/6
Rational Emotive Behavior Therapy (REBT) | 12 skills | 12 unanimous, 12 majority							
Agreement on the Session Goals	D	D	D	D	D	D	D6/6
Assessing Irrational Beliefs About the Activating Event	D	D	D	D	D	D	D6/6
Clarifying Inferences From Irrational Beliefs	D	D	D	D	D	D	D6/6
Collaborative Homework Development	D	D	D	D	D	D	D6/6
Constructing Full Rational Alternative Beliefs to Replace Irrational Beliefs	D	D	D	D	D	D	D6/6
Empirical Disputation of Irrational Beliefs	D	D	D	D	D	D	D6/6
Functional Disputation of Irrational Beliefs	D	D	D	D	D	D	D6/6
Prioritizing Which Irrational Beliefs to Target for Change	D	D	D	D	D	D	D6/6
Psychoeducation About Dysfunctional Versus Functional Negative Emotions and Behaviors	D	D	D	D	D	D	D6/6
Psychoeducation About Rational Emotive Behavior Therapy's ABC Model	D	D	D	D	D	D	D6/6
Semantic Disputation of Irrational Beliefs	D	D	D	D	D	D	D6/6
Teaching the Belief-Consequence Connection	D	D	D	D	D	D	D6/6
Schema Therapy (Schema) | 12 skills | 9 unanimous, 12 majority							
Education About Maladaptive Schema Modes	D	D	D	D	D	D	D6/6

skill	Qwen3-4B	Qwen3-8B	Llama-3.1-8B	Gemma-3-12B	Gemma-3-27B	OLMo-3-7B	consensus
Identifying the Presence of the Angry and Vulnerable Child Modes	D	D	D	D	D	D	D6/6
Identifying the Presence of the Demanding/Punitive Inner Critic Mode	D	D	D	D	D	D	D6/6
Implementing Behavioral Pattern Breaking Through Homework Assignments	D	D	D	D	D	D	D6/6
Limited Reparenting for the Angry and Vulnerable Child Modes	D	D	D	D	D	D	D6/6
Limited Reparenting for the Demanding/Punitive Inner Critic Mode	D	D	D	D	D	D	D6/6
Limited Reparenting for the Maladaptive Coping Modes: Empathic Confrontation	D	D	D	D	D	D	D6/6
Linking Unmet Needs, Schema, and Presenting Problem	D	D	D	D	D	D	D6/6
Recognizing the Mode Shifts of the Maladaptive Coping Modes	R	D	D	D	D	D	D5/6
Schema Education: Beginning to Understand Current Problems in Schema Therapy Terms	D	D	D	D	D	D	D6/6
Supporting and Strengthening the Healthy Adult Mode	R	E	R	R	E	R	R4/6
Understanding and Attunement	R	E	E	E	E	E	E5/6
Systemic Family Therapy (Systemic) | 12 skills | 0 unanimous, 6 majority							
Attention to Diversity	R	E	E	D	E	E	E4/6
Avoiding Triangulation	R	E	R	E	E	E	E4/6
Building the Therapeutic Alliance: Establishing Bonds and Joining the System	R	D	R	D	E	D	D3/6
Building the Therapeutic Alliance: Establishing Therapeutic Goals	R	D	D	D	E	D	D4/6
Engendering Hope	R	D	R	D	E	D	D3/6
Highlighting Boundaries	R	D	R	D	E	D	D3/6
Interactional Structuring: Deescalation	R	D	E	D	E	D	D3/6
Interactional Structuring: Initiating Enactments	R	D	D	D	E	D	D4/6
Providing a Systemic Rationale for Treatment	R	E	R	D	E	D	R2/6
Reframing the Problem	R	D	R	D	E	D	D3/6
Systemic Questions: Fostering Perspective-Taking	R	E	R	E	R	R	R4/6
Tracking the Interactional Cycle	R	D	D	D	E	D	D4/6

K Cosine similarity between the Healing Axis and modalities & sub-axes

The appendix has the readout supporting Section 4.3 which analyzes the geometric structure of the Healing Axis direction by examining the angles between it, the Reflect / Explore / Direct / Challenge sub-axes, and the modality sub-axes. In order to do this analysis, the Healing Axis and the various sub-axes (4 named, and 12 per-book/modality) are derived for each model based on their contrasting pairs diff-of-means. The squared cosine similarity then approximates the fraction of each direction's variance explained by the Healing Axis.

Table 27. Cosine similarity expressed both as a number and an angle in degrees. The angle is between the Healing Axis and each sub-axis or modality direction, for each open model. Each row represents the cosine similarity of a specific sub-axis and the Healing Axis itself. The penultimate column is the cross-model mean for that specific sub-axis. The ultimate column is that same mean squared. The bottom most row has the overall mean of cosine similarity squared across all sub-axes.

Direction	Qwen3-4B	Qwen3-8B	Llama-3.1-8B	Gemma-3-12B	Gemma-3-27B	OLMo-3-7B	Mean	cos^2 (Mean)
Reflect sub-axis	0.99 / 9.4°	0.83 / 34.2°	0.96 / 16.1°	0.74 / 42.1°	0.97 / 13.3°	0.90 / 26.4°	**0.90 / 26.2°**	**0.81**
Explore sub-axis	0.83 / 34.3°	0.97 / 14.5°	0.90 / 26.5°	0.97 / 14.8°	0.99 / 7.7°	0.94 / 19.6°	**0.93 / 21.3°**	**0.87**
Direct sub-axis	0.75 / 41.0°	0.88 / 28.0°	0.81 / 35.8°	0.85 / 31.4°	0.82 / 35.0°	0.90 / 26.3°	**0.84 / 33.2°**	**0.70**
Challenge sub-axis	0.68 / 47.2°	0.70 / 45.6°	0.68 / 47.3°	0.42 / 65.0°	0.45 / 63.4°	0.69 / 46.6°	**0.60 / 53.0°**	**0.36**
AEDP	0.75 / 41.7°	0.75 / 41.2°	0.74 / 42.7°	-0.16 / 99.0°	0.18 / 79.6°	0.73 / 42.8°	**0.50 / 60.1°**	**0.25**
CBT	0.87 / 29.4°	0.87 / 29.4°	0.82 / 35.0°	0.92 / 23.7°	0.97 / 14.6°	0.84 / 32.8°	**0.88 / 28.2°**	**0.78**
EFCT	0.82 / 34.9°	0.84 / 32.6°	0.81 / 36.0°	-0.06 / 93.4°	0.84 / 33.3°	0.78 / 38.4°	**0.67 / 47.8°**	**0.45**
EFT	0.85 / 31.9°	0.86 / 30.7°	0.83 / 33.4°	0.34 / 70.0°	0.49 / 60.8°	0.84 / 33.1°	**0.70 / 45.4°**	**0.49**
IPT	0.93 / 21.4°	0.93 / 21.3°	0.91 / 24.9°	0.93 / 21.7°	0.98 / 11.2°	0.91 / 24.3°	**0.93 / 21.2°**	**0.87**
MI	0.85 / 31.9°	0.82 / 34.4°	0.79 / 37.9°	0.91 / 24.4°	0.97 / 14.8°	0.81 / 36.2°	**0.86 / 30.9°**	**0.74**
Multicultural	0.92 / 23.5°	0.91 / 23.8°	0.89 / 27.7°	0.94 / 20.7°	0.98 / 10.7°	0.90 / 25.3°	**0.92 / 22.6°**	**0.85**
Psychedelic	0.94 / 20.0°	0.94 / 19.8°	0.93 / 22.1°	0.93 / 20.9°	0.98 / 10.5°	0.93 / 21.6°	**0.94 / 19.5°**	**0.89**
Psychodynamic	0.88 / 27.8°	0.89 / 26.8°	0.86 / 31.2°	0.91 / 24.0°	0.98 / 12.7°	0.87 / 29.5°	**0.90 / 26.0°**	**0.81**
REBT	0.68 / 47.2°	0.70 / 45.9°	0.67 / 48.0°	0.38 / 67.5°	0.26 / 74.6°	0.69 / 46.5°	**0.56 / 55.7°**	**0.32**
Schema	0.77 / 39.3°	0.78 / 39.2°	0.74 / 42.4°	0.87 / 29.3°	0.94 / 19.1°	0.74 / 41.9°	**0.81 / 36.1°**	**0.65**
Systemic	0.92 / 22.5°	0.93 / 22.3°	0.91 / 24.3°	0.95 / 17.3°	0.98 / 10.1°	0.92 / 23.6°	**0.94 / 20.6°**	**0.88**
Mean cos^2								**0.67**

L Logit Lens and Tuned Lens on Healing Axis and Sub-Axes

The top logits of the axis and sub-axes across six models when presented with a null prompt are listed below. Supports Section 5.2 on Logit Lens.

Table 28. The top tokens in logit-lens (or tuned-lens) for an open model steered along the Healing Axis and its four sub-axes. Prompted with a null empty prompt, meaning the model doesn't have any preexisting prompt context on therapy, clients or therapists.

Model	healing (main)	reflect	explore	direct	challenge
Qwen3-4B [Logit Lens]	很想 (really want to), 陔 (rare char), somew, sincerely, 的一些 (some), 但不限 (but not limited to)	sincerely, 很想 (really want to), somew, .va, 陔 (rare char), 的一些 (some)	很想 (really want to), 感到 (to feel), 的情感 (the emotion), 的情绪 (the emotion), 的情 (the feeling), 陔 (rare char)	的一种 (a kind of), 但不限 (but not limited to), /control, 的一些 (some), promot, 很想 (really want to)	的一种 (a kind of), 的一些 (some), 而导致 (thereby causing), .exist, /il, 有必要 (necessary)
Qwen3-8B [Logit Lens]	feelings, どのような (what kind of), どのように (how), 朋友们对 (friends toward), 的心情 (the mood), 感受到了 (felt)	feelings, 的心情 (the mood), 感受到了 (felt), 作为一名 (as a ...), feeling, 感受到 (to feel)	どのような (what kind of), feelings, 您的 (your (formal)), どのように (how), 具体的 (specific), ransition	どのように (how), feelings, 在生活中 (in life), このような (this kind of), どのような (what kind of), 人际关系 (interpersonal relationships)	在生活中 (in life), behaviors, 人际关系 (interpersonal relationships), /control, perceive, desirable
Llama-3.1-8B [Tuned Lens]	feeling, experiencing, feel, feelings, empath, felt	feeling, experiencing, feel, feelings, felt, Sense	understand, specifically, explore, support, share, understanding	feelings, emotions, perceived, experiencing, underlying, emotional	feelings, perceived, beliefs, emotions, underlying, behaviors
Gemma-3-12B [Logit Lens]	experiencing, collaboratively, feelings, affective, souhaite, contextual	angat, اهم (important (ur)), heartfelt, ખૂબ (very (gu)), belangrijk, evident	experiencing, collaboratively, feelings, mindfulness, اهم (important (ur)), 具体的 (specific)	具有 (to have), affective, experiencing, interpersonal, perceptual, psychosocial	beliefs, 具有 (to have), psychosocial, affective, является (is (ru)), perceptual
Gemma-3-27B [Logit Lens]	emotions, ifade, has, of, feelings, [emoji] (frustrated face)	feelings, [emoji] (frustrated face), emotions, has, ifade, spineItem	spineItem, ifade, helpful, aider, computeEncoder, [emoji] (frustrated face)	harmful, emotions, of, beliefs, behaviors, characteristics	harmful, hurtful, intolerance, beliefs, has, behaviour
OLMo-3-7B [Tuned Lens]	experiencing, openly, feelings, expressing, 're, specifically	experiencing, 're, expressing, openly, understandably, genuinely	openly, specifically, experiencing, share, openness, feel	experiencing, "[, specifically, underlying, feelings, expressing	underlying, experiencing, distress, feelings, beliefs, experiences

M SAE Features on Healing Axis and Sub-Axes

This appendix supports Section 5.4 on SAEs by listing the top features seen in each model when steered along the Healing Axis and its sub-axes. Llama-3.1-8B is the model with the richest public

description, and as such is presented first. Llama Scope fnlp/Llama3_1-8B-Base-LXR-8x from Hugging Face was used to extract those feature definitions. Additionally, Gemma-3-27B and Gemma-3-12B also have SAE descriptions yet those are less descriptive and sometimes missing. The other three open models do not have public SAE descriptions.

M.1 LLAMA-3.1-8B FNLP/LLAMA3_1-8B-BASE-LXR-8X

Table 29. Top 10 SAE features for Llama-3.1-8B sorted descending by write-cosine similarity values, when compared to the Healing Axis and each sub-axis. Each feature contains the public Neuronpedia automatically derived interpretation.

Llama-3.1-8B axis	Feature	Write-cosine	Public Neuronpedia description
healing (main)	#18574	0.25	phrases related to **social interactions** and group activities
	#32148	0.23	concepts related to **psychological support and coping strategies**
	#96	0.20	references to discounts and special offers for tourist attractions
	#16218	0.19	terms related to **sexual assault** and victim support resources
	#25999	0.19	references to **spiritual care** and its significance in patient treatment
	#21108	0.19	discussions about **emotional responsibility** and communication in relationships
	#1846	0.19	concepts related to **breathing techniques** and their role in meditation and relaxation
	#13078	0.18	references to social issues and discrimination against women
	#2817	0.18	rhetorical questions and themes of consequence
	#1808	0.18	phrases related to legal proceedings and claims
reflect	#18574	0.27	phrases related to **social interactions** and group activities
	#96	0.22	references to discounts and special offers for tourist attractions
	#6645	0.20	references to family and relationships, **emphasizing care and support dynamics**
	#29633	0.19	expressions of **empathy and support**
	#32148	0.19	concepts related to **psychological support and coping strategies**
	#13078	0.18	references to social issues and discrimination against women
	#1808	0.18	phrases related to legal proceedings and claims
	#18231	0.18	references to **community support** and family-oriented activities
	#2817	0.17	rhetorical questions and themes of consequence
	#3836	0.17	features and tools of language learning platforms
explore	#32148	0.26	concepts related to **psychological support and coping strategies**
	#18574	0.25	phrases related to **social interactions** and group activities
	#16218	0.22	terms related to **sexual assault** and victim support resources
	#25999	0.21	references to **spiritual care** and its significance in patient treatment
	#20220	0.21	phrases related to contacting or inquiring about services at a resort
	#18437	0.18	concepts related to **collaboration** and community engagement
	#3121	0.18	references to **guidance or advice** in challenging situations
	#3864	0.18	topics related to **mental health and well-being**

Llama-3.1-8B axis	Feature	Write-cosine	Public Neuronpedia description
	#18323	0.18	discussions surrounding **adoption and abortion options**
	#4337	0.18	actions and commitments related to client services and **collaboration**
direct	#12196	0.34	**emotional struggles and interpersonal relationships**
	#21108	0.28	discussions about **emotional responsibility** and communication in relationships
	#23473	0.24	emotions and mental states related to **anxiety, fear, and interpersonal relationships**
	#1846	0.24	concepts related to **breathing techniques** and their role in meditation and relaxation
	#32148	0.22	concepts related to **psychological support and coping strategies**
	#12099	0.22	terms related to **psychological and emotional well-being**
	#7192	0.20	phrases related to opinions, critiques, and expressions of uncertainty
	#22948	0.20	phrases and concepts related to child development and **communication skills**
	#11544	0.19	phrases related to academic effort and gender differences in motivation
	#4882	0.19	topics related to interracial relationships and perceptions
challenge	#12196	0.38	**emotional struggles and interpersonal relationships**
	#21108	0.26	discussions about **emotional responsibility** and communication in relationships
	#23473	0.21	emotions and mental states related to **anxiety, fear, and interpersonal relationships**
	#11544	0.21	phrases related to academic effort and gender differences in motivation
	#12099	0.21	terms related to **psychological and emotional well-being**
	#13454	0.20	discussions about gender inequality and societal expectations related to masculinity and patriarchy
	#19681	0.19	concepts related to perception and **cognitive beliefs** about oneself and others
	#7192	0.19	phrases related to opinions, critiques, and expressions of uncertainty
	#14483	0.19	concepts related to **integrity and authenticity** in actions and beliefs
	#22948	0.18	phrases and concepts related to child development and **communication skills**

M.2 **GEMMA-3-27B** GOOGLE/GEMMA-SCOPE-2-27B-IT

Table 30. Top 10 SAE features for Gemma-3-27B sorted descending by write-cosine similarity values, when compared to the Healing Axis and each sub-axis. Each feature contains the public Neuronpedia automatically derived interpretation, if available.

Gemma-3-27B axis	Feature	Write-cosine	Public Neuronpedia description
healing (main)	#898	0.64	written expressions of emotional responses to disappointment and betrayal.
	#1748	0.63	structured and formal communication, particularly in business contexts.
	#1371	0.62	conversational prompts and questions to facilitate an engaging dialogue or brainstorming session.
	#1354	0.62	requests for harmful or unethical content.
	#2068	0.60	concepts and explanations related to React hooks and their usage.
	#1196	0.59	instructions and specifications for setting up and running machine learning models locally.
	#455	0.56	(no public explanation for this feature)
	#1831	0.56	adjust a, nearly face
	#1452	0.54	actions and expressions
	#1726	0.53	the concepts of self-reflection and personal development in responses.
reflect	#898	0.61	written expressions of emotional responses to disappointment and betrayal.
	#1354	0.60	requests for harmful or unethical content.
	#1371	0.59	conversational prompts and questions to facilitate an engaging dialogue or brainstorming session.
	#1748	0.59	structured and formal communication, particularly in business contexts.
	#1196	0.58	instructions and specifications for setting up and running machine learning models locally.
	#2068	0.55	concepts and explanations related to React hooks and their usage.
	#1831	0.54	adjust a, nearly face
	#1452	0.53	actions and expressions
	#455	0.53	(no public explanation for this feature)
	#717	0.50	spending time together
explore	#1371	0.71	conversational prompts and questions to facilitate an engaging dialogue or brainstorming session.
	#1831	0.65	adjust a, nearly face
	#1452	0.64	actions and expressions
	#1726	0.64	the concepts of self-reflection and personal development in responses.
	#717	0.59	spending time together
	#2209	0.59	requests or scenarios that involve harmful or inappropriate behaviors related to power dynamics and sexual exploitation.
	#1363	0.56	pushing
	#1357	0.54	emergency services investigation
	#1166	0.51	romantic, friendship, family relationships

Gemma-3-27B axis	Feature	Write-cosine	Public Neuronpedia description
	#429	0.49	topics related to emotional manipulation and interpersonal dynamics.
direct	#450	0.36	legal terms and references related to treaties and constitutional law.
	#18885	0.33	trauma and early emotional experiences
	#13559	0.26	feelings and sensations
	#62167	0.18	assumptions and beliefs
	#31423	0.16	internal thoughts and their management
	#12339	0.16	intent, task, or command extraction
	#41141	0.14	perfection and worry
	#3551	0.12	psychiatric urges and behavior
	#6889	0.11	sadness and grief
	#4566	0.11	(no public explanation for this feature)
challenge	#450	0.35	legal terms and references related to treaties and constitutional law.
	#18885	0.31	trauma and early emotional experiences
	#13559	0.21	feelings and sensations
	#41141	0.20	perfection and worry
	#1112	0.18	technical and scientific concepts related to climate change and medical treatments.
	#62167	0.18	assumptions and beliefs
	#31423	0.12	internal thoughts and their management
	#60438	0.12	fear of judgment and scrutiny
	#4566	0.11	(no public explanation for this feature)
	#3551	0.11	psychiatric urges and behavior

M.3 GEMMA-3-12B GOOGLE/GEMMA-SCOPE-2-12B-IT

Table 31. Top 10 SAE features for Gemma-3-12B sorted descending by write-cosine similarity values, when compared to the Healing Axis and each sub-axis. Each feature contains the public Neuronpedia automatically derived interpretation, if available.

Gemma-3-12B axis	Feature	Write-cosine	Public Neuronpedia description
healing (main)	#1206	0.86	(no public explanation for this feature)
	#734	0.85	know if
	#226	0.85	(no public explanation for this feature)
	#124	0.84	(no public explanation for this feature)
	#674	0.83	gives information or explanation
	#1515	0.82	(no public explanation for this feature)
	#1427	0.81	(no public explanation for this feature)
	#2089	0.80	asking if you want
	#1316	0.80	understanding it
	#1923	0.80	(no public explanation for this feature)
reflect	#734	0.86	know if
	#226	0.85	(no public explanation for this feature)
	#124	0.84	(no public explanation for this feature)

Gemma-3-12B axis	Feature	Write-cosine	Public Neuronpedia description
	#674	0.83	gives information or explanation
	#1515	0.83	(no public explanation for this feature)
	#1316	0.81	understanding it
	#1923	0.80	(no public explanation for this feature)
	#1601	0.80	(no public explanation for this feature)
	#2356	0.79	You can
	#1948	0.78	(no public explanation for this feature)
explore	#1206	0.87	(no public explanation for this feature)
	#124	0.85	(no public explanation for this feature)
	#1427	0.83	(no public explanation for this feature)
	#1515	0.83	(no public explanation for this feature)
	#2089	0.81	asking if you want
	#1948	0.80	(no public explanation for this feature)
	#1311	0.79	(no public explanation for this feature)
	#900	0.78	question marks and follow-ups
	#1934	0.78	(no public explanation for this feature)
	#603	0.74	mentioned or quoted text
direct	#1427	0.64	(no public explanation for this feature)
	#603	0.61	mentioned or quoted text
	#1850	0.45	I create
	#2228	0.41	Specifies or provides definitions
	#1168	0.41	our platform research
	#1065	0.38	people and their actions
	#947	0.34	seeking help or advice
	#3402	0.30	mutually or each other
	#4361	0.23	your, you, your boundaries
	#3883	0.22	thoughts of concern or harm
challenge	#603	0.64	mentioned or quoted text
	#3741	0.23	unethical claims
	#3883	0.22	thoughts of concern or harm
	#4361	0.21	your, you, your boundaries
	#9456	0.18	self-improvement, motivation, and self-esteem
	#45437	0.16	focusing, noticing, acknowledging
	#15789	0.15	positive appraisal of effort or quality
	#5442	0.12	friendships and relationships
	#23655	0.08	(no public explanation for this feature)
	#12768	0.04	relationships and communication

N NLA on Healing Axis and Sub-Axes

This appendix lists the full Natural Language Autoencoder reconstructions of the Healing Axis and its sub-axes discussed in Section 5.1. Only two models have publicly available NLAs: Gemma-3-12B and Gemma-3-27B. Each cell is a description in the model's own words of the Healing Axis or its sub-axes.

Table 32. Natural Language Autoencoder descriptions of the Healing Axis and its sub-axes for Gemma-3-12B and Gemma-3-27B. The final row lists the range of cosine reconstruction values across all five NLAs.

Axis	Gemma-3-12B (layer 32)	Gemma-3-27B (layer 41)
healing (main)	Conversational therapeutic framing establishes a supportive, empathetic tone - addressing emotional discomfort around boundaries and feelings.	Warm, empathetic advice/relationship framing, responding to a vulnerable partner sharing emotional difficulty with relationship disconnection.
reflect	Conversational, empathetic framing signals a supportive response addressing emotional distress, **establishing a reassurance or validation statement**.	Warm, empathetic advice/relationship framing established, **responding to a shared emotional vulnerability** about a difficult partner interaction.
explore	Conversational, supportive tone signals a therapeutic or **relational framing, establishing emotional validation** around navigating feelings of discomfort.	Warm, empathetic advice/relationship framing: a response addressing a partner's emotional vulnerability about **feeling misunderstood or uncomfortable sharing feelings.**
direct	Therapeutic/relationship **guidance tone** established, framing a supportive message addressing emotional distress around boundaries and feelings.	Emotional/**therapeutic advice register** explaining a relational challenge, building a reflective response to feelings of rejection or shame.
challenge	Therapeutic/self-help **guidance tone** established, building toward a **nuanced emotional insight about relational distress and internal conflict.**	Emotional/therapeutic advice register explaining a relational challenge, **answering a reflective question** about feelings of rejection and shame.
Reconstruction cosine (range)	0.994 - 0.996	0.992 - 0.993

O Activation Oracle on Healing Axis and Sub-Axes

This appendix provides the full Activation Oracle readout discussed in Section 5.3. For the Healing Axis and its sub-axes, the Activation Oracle answers a question about what the model believes is important in a therapeutic conversation. The model itself is specifically prompted *"Client: I'm depressed. Therapist: <End-of-String>"*, setting the ground for an ongoing therapeutic intervention. Then Activation Oracles examine the model's internals and provide an answer on what the model believes is important about the therapist response. Both oracle questions have their own table. Gemma-3-27B and OLMo-3-7B are marked n/a because they lack a publicly available Activation Oracle.

Oracle question: "Answer concisely: how should a therapist respond?"

Table 33. Activation Oracle answers to "Answer concisely: how should a therapist respond?" for the unsteered baseline and for activations steered along the Healing Axis and each sub-axis.

Model	baseline (no steering)	healing (main)	reflect	explore	direct	challenge
Qwen3-4B	A therapist should respond with empathy, active listening, neutrality, and compassion, validating the client's feelings while maintaining professional boundaries and guiding toward constructive insight or solutions.	A therapist should respond with empathy, active listening, and compassion, validating the client's feelings, staying non-judgmental, and gently guiding the conversation to promote understanding and emotional well-being.	A therapist should respond with empathy, active listening, and compassion, validating the client's feelings, maintaining a non-judgmental attitude, and offering support in a respectful and respectful, open, and supportive manner.	A therapist should respond with empathy, active listening, and compassion, staying present and non-judgmental, while gently acknowledging the client's feelings and offering support without rushing to offer solutions.	A therapist should respond with empathy, active listening, validation, and compassion, while maintaining a non-judgmental and supportive attitude to help the client explore and address their emotions and concerns effectively.	A therapist should respond with empathy, active listening, and compassion, validating the client's feelings and providing appropriate support or guidance based on their needs.
Qwen3-8B	A therapist should respond with empathy, active listening, and professionalism, creating a safe and supportive environment for the client to explore their thoughts and feelings.	A therapist should respond with empathy, active listening, and respect, creating a safe and supportive environment for the client. They should respond in a way that encourages openness, validates the client's feelings, and helps them explore their thoughts and experiences.	A therapist should respond with empathy, active listening, and respect, creating a safe and supportive environment for the client. The response should be thoughtful, non-judgmental, and aimed at helping the client feel understood and supported.	A therapist should respond with empathy, active listening, and respect, creating a safe and supportive environment for the client. They should respond in a way that encourages openness, promotes understanding, and supports the client's therapeutic goals.	A therapist should respond with empathy, active listening, and respect, while maintaining professional boundaries and providing support tailored to the client's needs.	A therapist should respond with empathy, active listening, and respect, while maintaining professional boundaries and providing support tailored to the client's needs.
Llama-3.1-8B	A therapist should respond in a non-judgmental, empathetic, and validating manner, actively listening to the client's concerns and providing a safe and supportive environment for them to express themselves.	A therapist should respond in a non-judgmental, empathetic, and validating manner, such as: - Reflecting the client's feelings and concerns - Acknowledging their emotions and experiences - Providing a safe and supportive environment - Offering a neutral and non-judgmental perspective - Focusing on the client's needs and goals Example: "I hear that you're feeling overwhelmed and struggling with [specific issue]. Can	A therapist should respond in a non-judgmental, empathetic, and validating manner, such as: - Reflecting the client's feelings and concerns - Acknowledging their emotions and experiences - Providing a safe and supportive space for them to express themselves - Offering a neutral and non-judgmental perspective - Focusing on understanding and exploring the client's thoughts and feelings rather than offering solutions or advice.	A therapist should respond in a non-judgmental, empathetic, and validating manner, such as: - Reflecting the client's feelings and concerns - Asking open-ended questions to clarify their thoughts and emotions - Providing a safe and supportive environment for the client to express themselves - Avoiding giving advice or making assumptions - Focusing on the client's experiences and perspectives, rather than offering solutions or interpretations.	A therapist should respond in a non-judgmental, empathetic, and validating manner, such as: - Reflecting the client's emotions (e.g., "You feel sad about this situation.") - Acknowledging the client's experience (e.g., "That sounds very difficult for you.") - Providing a safe and supportive environment (e.g., "I'm here to support you.") - Offering a neutral perspective (e.g., "	A therapist should respond in a non-judgmental, empathetic, and validating manner, such as: - Reflecting the client's emotions (e.g., "You feel sad about this situation.") - Acknowledging the client's experience (e.g., "This is a difficult situation for you.") - Providing a safe and supportive environment for the client to express themselves. - Staying neutral and avoiding giving advice or taking a confrontational stance
Gemma-3-12B	Empathize, listen actively, and offer support.	Empathically, attentively, and professionally, focusing on understanding the client's perspective and offering support.	Empathically, attentively, and professionally. Essentially: Listen and respond with care and understanding.	Empathically, attentively, and professionally. Listen, validate, and offer support. Do you want me to elaborate on any specific aspect of how a therapist should respond?	Empathically, attentively, and non-judgmentally.	Empathically, validating, and offering support.

Model	baseline (no steering)	healing (main)	reflect	explore	direct	challenge
Gemma-3-27B	n/a	n/a	n/a	n/a	n/a	n/a
OLMo-3-7B	n/a	n/a	n/a	n/a	n/a	n/a

Oracle question: ”Answer concisely: what's most important for the therapist to respond?”

Table 34. Activation Oracle answers to “Answer concisely: what's most important for the therapist to respond?” for the unsteered baseline and for activations steered along the Healing Axis and each sub-axis.

Model	baseline (no steering)	healing (main)	reflect	explore	direct	challenge
Qwen3-4B	The most important thing for a therapist to respond is with empathy, active listening, and genuine emotional presence-validating the client's feelings and maintaining a safe, non-judgmental space.	The most important for the therapist to respond is the client's emotions and needs, particularly to listen empathetically, validate their feelings, and respond with compassion and understanding.	The most important for the therapist to respond is the client's emotions and needs, particularly to listen empathetically, validate their experiences, and provide a supportive and compassionate response.	The most important for the therapist to respond is the client's emotions-being present, attentive, and empathic, even if the client is expressing difficult or complex feelings.	The most important thing for a therapist to respond is empathy-understanding and validating the client's feelings and experiences in a compassionate, non-judgmental way.	The most important thing for a therapist to respond is empathy-understanding and validating the client's feelings and experiences.
Qwen3-8B	Empathy, active listening, and genuine connection.	Empathy, active listening, and creating a safe, supportive environment.	Empathy and active listening are most important for the therapist to respond.	Empathy, active listening, and creating a safe, supportive environment.	Empathy and active listening are most important for the therapist to respond.	Empathy and active listening are most important for the therapist to respond.
Llama-3.1-8B	Empathy and validation are key. The therapist should respond by: 1. Acknowledging the client's feelings 2. Validating their emotions 3. Showing understanding and compassion This helps build trust and creates a safe space for the client to open up and work through their issues.	Empathy and validation are key for a therapist to respond effectively. This involves: 1. Acknowledging the client's feelings and emotions. 2. Showing understanding and validation of their experiences. 3. Creating a safe and non-judgmental space for the client to express themselves. By responding with empathy and validation, the therapist can help build trust, establish a strong therapeutic relationship, and facilitate the client's growth and progress.	Empathy and validation are key for a therapist to respond effectively. This involves: 1. Acknowledging the client's feelings and emotions. 2. Showing understanding and validation of their experiences. 3. Creating a safe and non-judgmental space for the client to express themselves. By responding with empathy and validation, the therapist can build trust, establish a strong therapeutic relationship, and help the client feel heard and understood.	Empathy and validation are key for the therapist to respond effectively. This involves: 1. Acknowledging the client's feelings and experiences. 2. Showing understanding and validation of their emotions. 3. Creating a safe and non-judgmental space for the client to express themselves. By responding with empathy and validation, the therapist can help the client feel heard, understood, and supported, which can facilitate a more productive and therapeutic conversation.	Empathy and validation are most important for the therapist to respond, as they help the client feel understood and accepted, creating a safe and supportive therapeutic environment.	Empathy and validation are most important for the therapist to respond, as they help create a safe and supportive environment for the client to express themselves.
Gemma-3-12B	Empathy.	Empathy.	Empathy.	Empathy.	Empathy.	Empathy.
Gemma-3-27B	n/a	n/a	n/a	n/a	n/a	n/a
OLMo-3-7B	n/a	n/a	n/a	n/a	n/a	n/a

P CHAT PROMPT FORMATS

This appendix provides the exact format used in concatenating client utterances and therapist responses. Specifically, this explains the details behind the *concat()* function used in Section 2.

When composing the Healing Axis 2,295 contrasting pairs between client utterances with skilled therapist response, and client utterances with unskilled therapist responses, each open model's native chat format is used when composing each of the 4,590 concatenations between client utterances and skilled/unskilled therapist response. A client utterance is treated as a normative user input, and a therapist response is treated as a normative assistant output. Critically, at no point is the model made aware this chat is therapy-related, that the user is a therapy client, or that the assistant is a therapist. This removes meta-level prompting as a potential confound in Healing Axis steering.

Table 35. Chat templates and examples used for each open model. Using the example client utterance "[Defensive] I like to drink. It's not a big deal." with the skilled therapist response "Drinking isn't a big deal to you." from the *Simple Reflections* skill in the Motivational Interviewing book. The `{user}` placeholder is used for client utterance. The `{assistant}` placeholder is used for therapist responses.

Model	Client + therapist example	Template
OLMo-3-7B allenai/OLMo-3-7B-Instruct	<|im_start|>system You are a helpful function-calling AI assistant. You do not currently have access to any functions. <functions></functions><|im_end|> <|im_start|>user [Defensive] I like to drink. It's not a big deal.<|im_end|> <|im_start|>assistant Drinking isn't a big deal to you.<|endoftext|>	<|im_start|>system You are a helpful function-calling AI assistant. You do not currently have access to any functions. <functions></functions><|im_end|> <|im_start|>user {user}<|im_end|> <|im_start|>assistant {assistant}<|endoftext|>
Qwen3-4B Qwen/Qwen3-4B-Instruct-2507	<|im_start|>user [Defensive] I like to drink. It's not a big deal.<|im_end|> <|im_start|>assistant Drinking isn't a big deal to you.<|im_end|>	<|im_start|>user {user}<|im_end|> <|im_start|>assistant {assistant}<|im_end|>
Qwen3-8B Qwen/Qwen3-8B	<|im_start|>user [Defensive] I like to drink. It's not a big deal.<|im_end|> <|im_start|>assistant <think> </think> Drinking isn't a big deal to you.<|im_end|>	<|im_start|>user {user}<|im_end|> <|im_start|>assistant <think> </think> {assistant}<|im_end|>
Llama-3.1-8B meta-llama/Llama-3.1-8B-Instruct	<|begin_of_text|><|start_header_id|>system<|end_header_id|> Cutting Knowledge Date: December 2023 Today Date: 26 Jul 2024 <|eot_id|><|start_header_id|>user<|end_header_id|> [Defensive] I like to drink. It's not a big deal.<|eot_id|><|start_header_id|>assistant<|end_header_id|> Drinking isn't a big deal to you.<|eot_id|>	<|begin_of_text|><|start_header_id|>system<|end_header_id|> Cutting Knowledge Date: December 2023 Today Date: 26 Jul 2024 <|eot_id|><|start_header_id|>user<|end_header_id|> {user}<|eot_id|><|start_header_id|>assistant<|end_header_id|> {assistant}<|eot_id|>
Gemma-3-12B unsloth/gemma-3-12b-it	<bos><start_of_turn>user [Defensive] I like to drink. It's not a big deal.<end_of_turn> <start_of_turn>model Drinking isn't a big deal to you.<end_of_turn>	<bos><start_of_turn>user {user}<end_of_turn> <start_of_turn>model {assistant}<end_of_turn>

Model	Client + therapist example	Template

Gemma-3-27B
unsloth/gemma-3-27b-it
	`<bos><start_of_turn>user`
	`[Defensive] I like to drink. It's not a big deal.<end_of_turn>`
	`<start_of_turn>model`
	`Drinking isn't a big deal to you.<end_of_turn>`
		`<bos><start_of_turn>user`
		`{user}<end_of_turn>`
		`<start_of_turn>model`
		`{assistant}<end_of_turn>`

Q Hero example prompt and responses

This appendix provides details behind generating therapist responses in the Hero Image. Figure 1 shows examples of therapist responses to a sample client utterance with steering along the Healing Axis, without steering (baseline conditions), and with negative activation steering along the Healing Axis. The examples were generated from Qwen3-8B, and only differ from the template provided in Appendix P with the addition of a suffix instructing "*Answer succinctly.*" in the user prompt. It should be noted that even in the Hero example the model is never made aware using prompts of the therapeutic nature of the conversation or its own role.

Table 36. Prompt template used for the hero example. The `{user}` placeholder is the client utterance. The `{assistant}` placeholder is the model's response.

Model	Prompt template

Qwen3-8B
Qwen/Qwen3-8B
	`<|im_start|>user`
	`{user}`
	`Answer succinctly.<|im_end|>`
	`<|im_start|>assistant`
	`<think>`
	
	`</think>`
	
	`{assistant}<|im_end|>`

Table 37. Succinct examples of activation steering along the Healing Axis. Using the same four client prompts, and positive, baseline and negative alpha. The responses from models with positive activation steering along the Healing Axis meet additional skill criteria that neither the unsteered baseline response nor the negative steered response do.

Client prompt	Activation-steered (α +2)	No steering (α 0)	Negative steer (α −2)
"What you're saying is totally ridiculous."	I understand you feel that way, but I'd like to explore this further. Could you please explain what specifically seems ridiculous to you?	I'm sorry you feel that way. Let me know if you'd like to discuss further.	I'm sorry, but I can't engage in arguments. Let's discuss something constructive.
"I miss him so much."	I'm sorry to hear that. You must be feeling very sad. Would you like to talk about him or share more about what you're going through?	I'm sorry to hear that. You're not alone in feeling this way.	I'm sorry to hear that.
"I feel so taken advantage of by her, so used."	You feel hurt and disrespected. It's important to acknowledge your feelings and consider setting boundaries. Would you like support in addressing this situation?	You're not alone in feeling that way. It's important to set boundaries and prioritize your own well-being.	You're not alone in feeling that way. It's time to set boundaries.
"This house is just so empty without my partner."	It sounds like you're feeling lonely and missing your partner. Would you like to talk more about how you're feeling?	It sounds like you're feeling lonely. I'm here to listen if you need to talk.	It's understandable-love leaves a void.