arXiv:2607.02885v1 [cs.CL] 3 Jul 2026

# Where do LLMs Fall Short in CBT-Guided Affective Reasoning?

**Vaishnavi Sinha[1]*, Pooja Guttal[1]*, Pranay Deep Reddy Katike[1], Vishal Sinha[4], Gerald Ndawula[1], Lira Yoon[2], Andrea Kleinsmith[3], and Manas Gaur[1]**
{vsinha1, poojag3, pranayk2, geraldn1, lyoon, andreak, manas}@umbc.edu, sinhav2@uci.edu
Departments of [1]Computer Science and Electrical Engineering, [2]Psychology, [3]Information Systems, UMBC
[4]Department of Electrical Engineering and Computer Science, UC Irvine

***Abstract*—Cognitive Behavioral Therapy (CBT) provides a structured framework for understanding a user's mental state by examining the interaction between cognitive and behavioral factors. However, out-of-the-box LLMs respond fluently and empathetically, yet collapse into validation & reflection, regardless of what the user actually needs. They know theoretical CBT (scoring up to 96% accuracy on licensing exam questions) but fail to apply it effectively. We explore this gap with a knowledge-guided framework that treats CBT dialogue as controlled affective reasoning: user narratives are decomposed into Beck's Cognitive Conceptualization structure, grounded in clinical SNOMED CT concepts validated via Natural Language Inference, and a Multiple Chain-of-Thought (MCoT) strategy selection between Validation & Reflection, Socratic Questioning, or Alternative Perspectives. To measure whether such guidance actually changes behavior, we introduce the Protocol Leverage Force ($\mathcal{F}$), a behavior-level metric that captures how far an intervention shifts a model away from its default response. Across three open-weight LLMs and 14 RealCBT-derived case studies, evaluated with human experts, valence–arousal trajectories, and linguistic entrainment, $\mathcal{F}$ shows that simply introducing protocol definitions via single chain-of-thought prompting fails to change LLM behavior, while MCoT on these definitions guides strategy selection better. Still, the effect stays within 1% (∼1.2%–1.3%), and all models remain biased toward Validation & Reflection. These results show CBT knowledge alone does not ensure effective application, giving the affective-computing community instrumentation to measure where LLMs fall short. Code repository: https://github.com/cbt-llm/cbt-llm***

***Index Terms*—Cognitive behavioral therapy, large language models, clinical knowledge grounding, mental health, valence–arousal, linguistic entrainment**

## I. Introduction

Millions now turn to AI for emotional support. A 2025 survey found therapy to be the leading real-world use case of generative AI [1], [2], and ChatGPT reports over 800 million weekly users as of late 2025, with average session lengths exceeding ten minutes [3], [4]. These are not casual exchanges but high-stakes conversations unfolding at enormous scale. Picard long ago argued that intelligent systems should recognize, understand, and express human emotion [5]; today's usage patterns suggest that vision has quietly arrived; LLMs have become a default medium for seeking emotional support, precisely because their responses feel natural, empathetic, and emotionally attuned [6]. Yet users bring real distress and genuine psychological struggle to systems with no structured way to unpack that distress or respond to it in a manner grounded in therapeutic practice. LLMs may sound emotionally aware [7], but fluency is not the same as care: we still cannot say which strategy a model is actually using, whether it is tracking the user's shifting emotional state, or whether the conversation leaves the person better off.

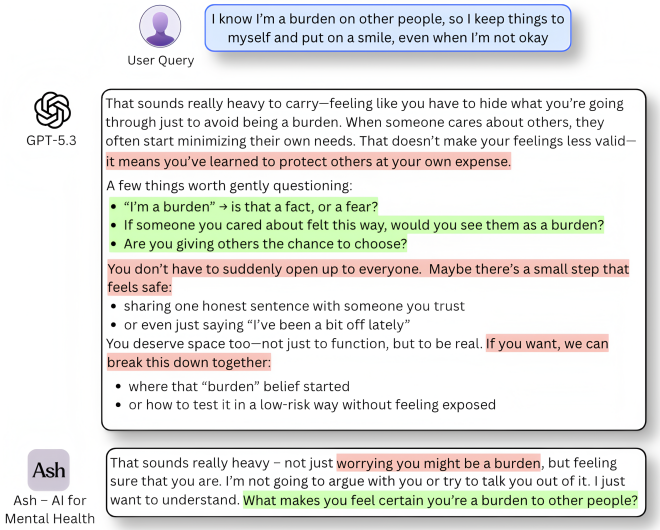



Fig. 1: Current response strategies of an out-of-the-box LLM, `GPT-5.3` (top), and `Ash`—AI for Mental Health (bottom) given the same initial user query without prior conversational context. `GPT-5.3` makes early assumptions (*"It just means you've learned to protect others"*) and moves towards premature problem-solving (*"we can break this down together"*). Whereas, `Ash` prioritizes questioning early on (*"What makes you feel certain"*) to elicit the user's cognitive context.

The gold standard for structured psychological intervention remains Cognitive Behavioral Therapy (CBT) [8], grounded in Beck's cognitive model. Distress arises not from situations themselves but from how individuals interpret them [9]. A CBT therapist first helps the client decompose experience into automatic thoughts, emotional responses, and behavioral consequences. The therapist then chooses an intervention based on where in that structure the unhelpful thought sits and how the person feels in the moment. A system that skips straight

*equal contribution.

to an empathetic response, without building this structure first, bypasses the very formulation that makes CBT work.

Out-of-the-box LLMs betray this gap in measurable ways. They exhibit a systematic *appraisal bias*, defaulting to a low-power, high-agreeableness stance no matter the context [7]. They produce dialogue with less entrainment than even non-expert peer supporters manage [10]. Their emotional arcs diverge systematically from those of real CBT sessions [11]. Prior work has made real progress adjacent to this problem, benchmarking CBT sub-skills [12], grounding patient simulation in Beck's model [13], and targeting cognitive restructuring directly [14]. But none of it grounds intervention selection in the user's own decomposed cognitive state, nor does any ask whether the resulting emotional trajectory resembles an authentic therapeutic interaction. The gap, then, is not one of response quality; rather, it is a gap in *principled affective reasoning*. We build a knowledge-guided framework that treats therapeutic dialogue as a controlled affective reasoning process, exposing where LLMs fall short of CBT principles. We make the following contributions.

- **Clinical grounding with inference verification:** We map user narratives to mental health findings in SNOMED CT (**S**ystematized **No**menclature of **Med**icine – **C**linical **T**erms), a clinically verified ontology, and verify these mappings through entailment-based inference. This verification propagates to downstream reasoning, grounding affective interpretation in clinically validated concepts rather than surface-level language.
- **Appraisal-based cognitive model and CBT-principle selection:** Drawing on Beck's Cognitive Conceptualization Diagram (CCD) [9], we decompose user queries into triggers, automatic thoughts, emotions, and behaviors. This decomposition exposes the user's cognitive model to the LLM, constraining its downstream reasoning to one of three CBT principles: Validation & Reflection (V), Socratic Questioning (SQ), or Alternative Perspectives (AP).
- **Affective trajectory evaluation:** We compare the valence and arousal [15], [16] of synthetic users against RealCBT clients, revealing how emotions in simulated scenarios diverge from those in real clinical settings. We further compute entrainment scores across simulated transcripts to measure synthetic user–therapist LLM synchrony [10], [17].
- **Protocol Leverage Force:** We introduce $\mathcal{F}$, a model-level behavior metric that quantifies how far our CBT-guided approach shifts a model's responses from its default baseline behavior.

## II. Related Work

**Cognitive Behavioral Therapy:** According to Beck, distress arises from how events are interpreted, not the events themselves. To better understand the relations, his CCD [9] decomposes each distress episode into triggering situations, automatic thoughts, emotional responses, and behavioral consequences [8]. Prior LLM work has engaged with parts of this structure without reasoning across the whole structure. CBT-Bench [12] evaluates competency across CBT sub-skills but measures isolated capabilities rather than reasoning from cognitive state to strategy. CRISP [14] targets cognitive restructuring as multi-turn dialogue but focuses on a single technique without modeling the appraisal state that determines when restructuring is warranted. PATIENT-Ψ [13] grounds patient simulation in Beck's model but addresses simulation rather than intervention selection. Kim et al. [18] align LLM behavior with therapist actions via instruction tuning but do not expose the CCD decomposition to the strategy selection mechanism at runtime. Our work treats the CCD decomposition itself as a knowledge input to response generation, grounding the LLM's strategy selection in the user's own surfaced cognitive structure.

**Sentiment Analysis with VAD Affective Dimensions** Beyond selecting the right CBT principle, we also need to verify whether the resulting conversation actually feels different to the user, which requires a way to track affect over time. Russell's circumplex model offers this by representing affect in a continuous valence-arousal space [15], capturing emotional experience in terms of pleasantness and intensity. An existing VAD lexicon [16] provides human-validated ratings for over 20,000 English words, enabling turn-level affective tracking without a trained classifier. Wang et al. [11] find that LLM-generated CBT sessions exhibit systematically different emotional arcs from human therapist sessions, particularly in valence recovery. We use their human-session data as the reference distribution against which we evaluate our own trajectories refined by knowledge-guided prompting.

**Linguistic Entrainment in Therapeutic Dialogue:** Valence and arousal trajectories capture *what* a user feels over the course of a session, but not *how closely* the two speakers' language stays aligned with one another, a property linguistic entrainment is designed to measure. *Entrainment* refers to speakers' tendency to align their language with one another as a conversation unfolds, an alignment known to build rapport and shown empirically to relate to therapeutic alliance and outcome [19], [20]. Kian et al. [10] apply normalized Conversational Linguistic Distance (nCLiD) [17], computing entrainment as the word-mover's distance between turns [21], to evaluate LLMs on CBT dialogue, finding that `GPT-3.5-turbo` significantly underperforms human therapists ($p < 0.001$). Delaherche et al. [19] caution that such metrics capture surface stylistic convergence but miss affective depth and temporal dynamics. We therefore pair nCLiD with valence–arousal trajectories, so that *surface-level alignment* and *affective trajectory* jointly inform our *assessment of dialogue quality*. We note that we use *entrainment* rather than *synchrony* [19], following its use in prior computational dialogue evaluation [10], [17].

## III. Building the CBT-guided Framework

### A. Dataset: Forming Synthetic User Case Studies

For our experiments, we extract 14 synthetic user case studies from the RealCBT dataset [11], a corpus of professional CBT therapist–client role-plays transcribed from public video

platforms. Leveraging its core issue labels to simulate real-world patient concerns, we sampled transcripts across 8 core issues: *career*, *self-esteem*, *health-related*, *anxiety*, *financial*, *academic*, *relationships*, and *other miscellaneous* concerns. Rather than reuse RealCBT's fixed client turns, which break conversational flow, we select a turn that captures the core issue, and simulate subsequent user turns by `GPT-4o-mini`. Since instruction-following and context-tracking degrade over extended interactions [22], we fixed our multi-turn sessions to 10 turns.

### B. Appraisal-Based User Cognitive Model.

Cognitive Appraisal Theory holds that emotional responses are determined not by events themselves but by how individuals interpret and evaluate them [23]. The same situation can produce fundamentally different emotional reactions across individuals, depending on their primary appraisal of its significance and their secondary appraisal of available coping resources. As emotions are shaped by interpretation rather than circumstance, they can be modified by changing how situations are appraised. Beck's CCD [9] characterizes the beliefs people hold and how those beliefs shape their feelings and actions. Within this framework, the Problem-Based Cognitive Conceptualization Diagram (PBCCD) provides a structured method for understanding a client's difficulties and identifying the most unhelpful cognitions and behaviors to target. For the example in Fig. 2, we ground our decomposition in this instrument, representing each user query as a *User Cognitive Model* that captures triggering situations, automatic thoughts, emotional responses, and behavioral consequences.

We implement this decomposition using `GPT-4o-mini` with a prompting scheme adapted from [13] and theorized with Beck's CCD [9]. For the running example, the decomposition yields: *Triggers* = $\emptyset$; *Automatic Thoughts* = "I'm a burden on other people"; *Emotions* = "shame, sadness"; *Behaviors* = "keep things to myself, put on a smile" (Fig. 2). Notably, the absence of an explicit trigger reflects insufficient context in the query to identify a discrete situational event; the automatic thought instead presents as a standing core belief rather than a situational reaction. The extracted behaviors further reveal a suppression pattern of self-concealment ("keep things to myself") paired with affective masking ("put on a smile"), consistent with expressive suppression documented in individuals experiencing negative affect [24], [25]. Making this latent cognitive structure explicit enables the system to identify the automatic thought as the target for intervention, without requiring the user to articulate it explicitly.

### C. SNOMED-CT Clinical Grounding with NLI Verification

Integrating clinical knowledge into language models improves semantic grounding and reduces hallucinations [26], [27], while explicit inference verification improves reasoning reliability by enforcing consistency between inputs and outputs [28]. Building on these insights, we ground language model inference in a structured clinical representation to enable clinically consistent affective reasoning.

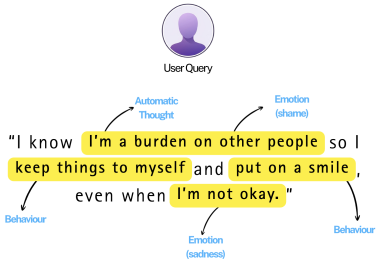



Fig. 2: **User Cognitive Model.** This cognitive structure enables explicit modeling of the user's affective state by linking triggers, thoughts, emotions, and behaviors, supporting more interpretable affective reasoning in downstream response generation.

We construct a graph-based affective representation from SNOMED CT [29] by extracting a focused subset of mental health–related concepts. Specifically, we build a subgraph centered on mental state and psychosocial functioning, capturing clinically relevant entities and their relationships.

The subgraph uses *IS_A* relations to model hierarchical abstraction from specific behaviors to higher-level traits, and *INTERPRETS* relations to link observable expressions with underlying psychological constructs. User narratives are mapped onto this representation, transforming subjective language into clinically grounded CBT-relevant concepts.

*1)* ***Clinical Concept Retrieval****:* To align natural language with clinical concepts, we embed [30] user utterances and SNOMED CT concepts in a shared semantic space. At inference time, we retrieve the top-$k = 5$ clinically relevant concepts using cosine similarity, balancing relevance coverage with noise reduction and interpretability [31], [32].

*2)* ***NLI-Based Clinical Concept Filtering****:* To ensure semantic coherence between user narratives and retrieved clinical concepts, we introduce a Natural Language Inference (NLI)-based verification layer. Similarity-based retrieval identifies related concepts but does not guarantee logical support from the user's state, risking incorrect clinical interpretations. As shown in Figure 3, each concept is reformulated as a hypothesis (e.g., "The user exhibits [clinical concept]") and evaluated against the user narrative using an NLI classifier, producing entailment, neutral, or contradiction labels. These labels enable contrastive reasoning over supported and conflicting interpretations. This replaces surface similarity with inference validation, improving reliability, interpretability, and clinical grounding of affective reasoning [33].

### D. CBT Principles

LLM response generations are grounded in three core CBT intervention principles: (1) *Validation & Reflection (V)*: Acknowledges the user's current emotions based on how they see the situation, without challenging, reframing, or reinforcing the underlying unhelpful thought; *(2) Socratic Questioning (SQ)*: Probes the user to help examine the thoughts and appraisals driving their distress; and *(3) Alternative Perspective (AP)*: Supports the user in constructing a more adaptive construal of the triggering situation by exploring counter-evidence, exceptions, or reframings of the user query.

### E. Knowledge-Guided Affective Reasoning

We operationalize three core CBT strategies (*V, SQ, AP*) as reasoning principles selected via **Multiple Chains-of-**

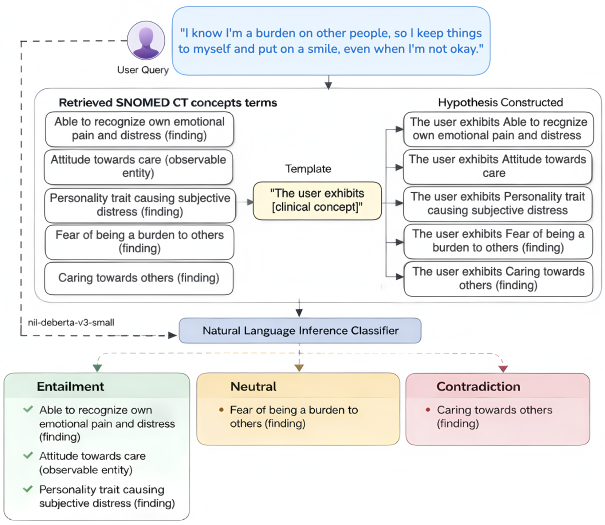



Fig. 3: Illustrating how a user's query is grounded in clinical terminologies through top-5 semantic retrieval and validated via contrastive NLI-based reasoning to categorize clinical concepts as Entailment, Neutral, or Contradiction based on the user's expressed narrative.

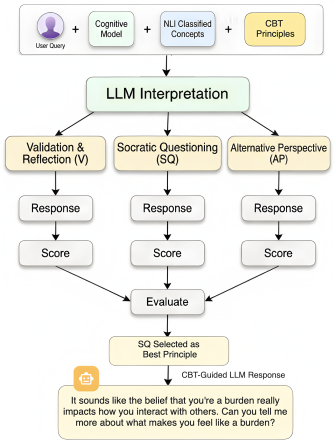



Fig. 4: MULTIPLE CHAIN-OF-THOUGHT REASONING. The model generates candidate responses for *(V), (SQ), and (AP)*, scores each candidate, and selects the highest-scoring response before outputting to the user.

**Thought (MCoT)** (Fig 4), which reasons over all three principles in parallel so each turn's strategy is observable and comparable against baseline. For each user turn, the LLM receives the user cognitive model (triggers, automatic thoughts, emotions, behaviors), NLI-classified SNOMED CT concepts (§III-C), and the guidelines per principle. It produces one candidate under each principle, each preceded by a reasoning trace, then scores them and selects the one that best advances the conversation. The explicit protocol labels let us analyze how often each strategy appears and how well the model's choices align with our evaluations. Temperatures were fixed for consistency: low (near-deterministic) for the therapist model, higher for the simulated user to allow variability. We evaluate across `Gemma3-12B`, `Mistral-7B`, and `GPT-OSS-20B` (hereafter referenced as `Gemma3`, `Mistral`, and `GPT-OSS`).

| Model | Q&A Acc. (%) |
|---|---|
| Mistral | 74.55 |
| Gemma3 | 80.45 |
| GPT-OSS | 89.86 |
| DeepSeek-R1-8B | 96.00 |

TABLE I: CBT-BENCH MCQ KNOWLEDGE TASK. Comparison of model performance on Q&A accuracy [12].

### F. CBT Theory vs. Application Gap

To investigate the disparity between theoretical knowledge and practical therapeutic application in LLMs, we benchmarked four models on the CBT-Bench Task 1 Q&A dataset [12], a 220-question multiple-choice evaluation derived from CBT licensing exams for MSW (Master of Social Work) students and compiled by domain experts. As shown in Table I, all models score between 74.55% and 96%, indicating high theoretical competency. Yet this competence does not translate to application (Appendix B): in baseline settings, models frequently default to *V* with assumptions not grounded in the user's cognitive context. `DeepSeek-R1-8B` tops theoretical accuracy (96%, Table I) but is weakest on applied measures: lowest human-rated effectiveness (3.53; the only statistically significant gap, p=0.002) and the most severe *(V)* bias (66.7% and 7.4% *(SQ)*) Table B.1. Given this misalignment, we exclude `DeepSeek-R1-8B` from our main evaluation (findings in Appendix B).

### G. Evaluation Criteria

*1)* ***Human Expert Evaluation****:* Eight domain-informed reviewers[1] blindly evaluated three transcripts each (one per model) through a Qualtrics survey. Each transcript was rated on two dimensions: (1) principle confidence, where annotators rated on a 0–4 scale how strongly each CBT principle *(V, SQ, AP)* appeared in the response, with a *Not Applicable* option; and (2) response preference, where annotators chose to either keep the final response, select one of the two alternative responses the model generated for the other CBT principles, or select *Neither*. We report inter-rater agreement measured using Krippendorff's $\alpha$ and Fleiss's $\kappa$, respectively.

*2)* ***Linguistic Entrainment****:* We evaluate transcripts via nCLiD [17], which measures how closely the LLM's turns track the user's language over a session. Utterances are encoded with `all-mpnet-base-v2` [34], giving 768-d contextual embeddings $\mathbf{e}(a_i)$ and $\mathbf{e}(c_j)$ for LLM turn $a_i$ and user turn $c_j$ respectively; these capture sentence-level semantics more robustly than the static word2vec representations used in [10], [17]. For each LLM turn $a_i$, we take the minimum cosine distance to any of the next $k$ user turns $c_j$, so that entrainment emerges across a short window, not just the adjacent turn is captured:

$$d_i^{C\to A} = \min_{i\le j\le i+k-1} 1 - \cos\big(\mathbf{e}(a_i), \mathbf{e}(c_j)\big)$$

where $i$ indexes therapist turns and $j$ ranges over user turns in the look-ahead window $[i, i+k)$, with $k = 2$. The window

[1]Evaluators comprised of two professors (background in clinical psychology and in CS/AI for mental health), one CS/psychology researcher, three graduate psychology students (MS/PhD), and two industry practitioners.

looks forward, following the nCLiD formulation [17], because a therapist's turn responds to what came before, so coordination may only surface later in the turn.

Transcript-level nCLiD averages these distances and normalizes by $\alpha$, the mean cosine distance across all turn pairs in the transcript, which corrects for within-speaker consistency and topic-driven similarity that would otherwise inflate entrainment estimates:

$$\mathbf{nCLiD} = \frac{\frac{1}{N}\sum_{i=1}^{N} d_i^{C\to A}}{\alpha} \tag{1}$$

with $N = 10$ turn pairs per transcript. Lower nCLiD indicates greater linguistic alignment with the user. We treat nCLiD as complementary to our valence–arousal trajectories, since semantic similarity captures linguistic convergence but not emotional tone or its shifts.

*3) **Protocol Leverage Force**:* As model deployments scale, evaluation increasingly requires surfacing behavior-level patterns that aggregate metrics cannot capture [35]. Following recent work showing that accuracy and F1 metrics can reward confident guessing over admitting uncertainty [36], we favor behavior-level metrics. Thus, instead of scoring correctness, we summarize the per-turn effect of the CBT-guided framework with a scalar we call *Protocol Leverage Force*, denoted $\mathcal{F}$, a measure of behavioral reorientation, not therapeutic quality. The name reflects what the quantity measures: the mechanical advantage of a structured intervention, i.e. how much response reorientation is achieved per unit of deviation from the model's native response manifold. Inspired by the "physics of language models" perspective that studies LLM behavior as a dynamical system [37], and following Huygens's classical centrifugal force $F = mv^2/r$ [38], we formally define $\mathcal{F}$ as follows:

$$\mathcal{F} = \frac{w_{\text{protocol}} \cdot \Delta^2}{d_{\cos}(e_{\text{MCoT}},\, \mu_{\text{base}}) + \varepsilon}, \tag{2}$$

$$\Delta = d_{\cos}(\text{baseline},\, \text{patient}) - d_{\cos}(\text{MCoT},\, \text{patient}),$$

where each factor captures a distinct dimension of the intervention: $\mathbf{w}_{\text{protocol}}$: the weight of the CBT principle *(V, SQ, AP)* selected on this turn. A confidently chosen protocol contributes more than an ambivalent one; this term reflects the framework's decision. $\boldsymbol{\Delta}$: the signed reorientation gain. Positive $\Delta$ means the MCoT response moved closer to the patient's surfaced cognitive state than the baseline would have. Squaring it ensures that the magnitude of reorientation dominates the signal while the sign is preserved separately for interpretation. $\mathbf{d}_{\cos}(\mathbf{e}_{\text{MCoT}}, \mu_{\text{base}}) + \varepsilon$: How far the MCoT response has strayed from the centroid of baseline responses in embedding space. $\varepsilon$ is the *baseline perturbation*: the mean distance between baseline centroids across three paraphrases of the baseline prompt, to test sensitivity to prompt wording.

$\mathcal{F}$ is large when a committed protocol produces a decisive reorientation toward the patient without the response drifting far from what the model would natively say. $\mathcal{F}$ is small under three distinct failure modes, each diagnosable from its terms: (i) low $w_{\text{protocol}}$: the framework was uncertain which strategy to apply; (ii) low $\Delta$: the response barely moved the dialogue relative to baseline; (iii) high denominator with moderate numerator: the model reframed stylistically but did not convert that distance into therapeutic gain. Reading the three terms jointly therefore exposes *why* a turn succeeded or failed, not only whether it did. We additionally examine $\Delta$ in isolation, and by applying Otsu's method [39] per model and CBT principle to avoid an arbitrary cutoff, separating high- and low-shift turns. These are reported in Appendix C and Table C.1.

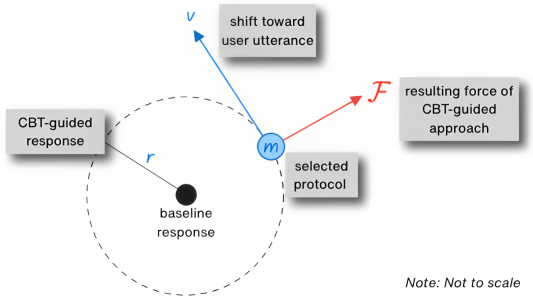



Fig. 5: Centrifugal-force-inspired view of $\mathcal{F}$, following Huygens's $F = mv^2/r$ [38]. The selected protocol is a point mass $m$ orbiting the baseline centroid (the "gravity" of the model's default behavior) at radius $r = d_{\cos}(\mathbf{e}_{\text{MCoT}}, \boldsymbol{\mu}_{\text{base}}) + \varepsilon$. The tangential arrow $v$ is the signed shift of the response toward the user's utterance; the outward arrow is the resulting force $\mathcal{F}$, which grows when the CBT-guided approach moves the response away from baseline and toward the user.

## IV. Results & Discussions

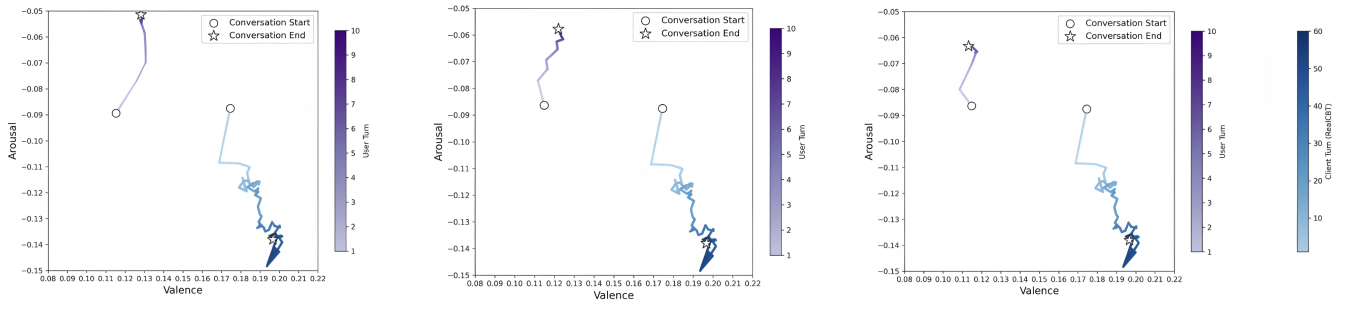
### A. Linguistic Entrainment & Valence–Arousal Trajectory

Across both synthetic and RealCBT sessions (Fig. 6), valence rises over time, marking a shift toward positive affect. Arousal tells a different story. RealCBT's 60-turn sessions de-escalate steadily, while our 10-turn synthetic sessions drift upward instead of calming. RealCBT's longer arcs are also richer, oscillating more than the synthetic ones. We read this divergence, not as a failure of the synthetic sessions to be lively, but as evidence that they capture the direction of valence change without the sustained de-escalation and affective depth of real therapy that Wang et al. document [11].

Table III reports nCLiD across models. `GPT-OSS` scores lowest, followed by `Gemma3` and `Mistral`, and *(V)* yields the highest entrainment in every model. Strikingly, our models entrain more than RealCBT does, the reverse of what Kian et al. found for `GPT-3.5-turbo`, which entrained less than human therapists [10]. The tension this reveals is telling. Synthetic sessions lean on *(V)*, which mirrors the user's language into smoother, tighter arcs, while RealCBT's oscillations likely stem from therapists' heavier use of *(SQ)*. In short, *(V)* maximizes surface entrainment, but *(SQ)* is what produces the affective arcs characteristic of real CBT.

### B. Linguistic Shift & Therapeutic Force

*1) **Validation and Socratic Questioning**:* Otsu-thresholded shift magnitudes and flagged counts (Appendix C, Table C.1) yield two rankings that, read together, separate CBT




(a) Mistral vs RealCBT (b) Gemma3 vs RealCBT (c) GPT-OSS vs RealCBT

Fig. 6: Valence–Arousal trajectories of user language across CBT-MCoT synthetic sessions (purple) and RealCBT human transcripts (blue), scored using the NRC VAD Lexicon. Each point is a cumulative mean over turns, and each graph shows the average arc per model across 14 transcripts. RealCBT's longer sessions reveal a more nuanced emotional journey (more oscillations).

TABLE II: Comparing baseline and MCoT examples for $\mathcal{F}$ (Eq. (2)). Both models receive the *same* user query ("I'm extremely good at my job. I mean, I'm the best employee that they have ever had... I make no errors. So how can I be wrong when it comes to these circumstances with my coworkers?"). Red marks early assumptions dropped in MCoT; green marks the new shift MCoT introduces. Orange marks nearly unchanged phrasing.

| Model/Mode | Response | $\mathcal{F}$ |
|---|---|---|
| Gemma3 *Baseline* | It's fantastic you're so confident and skilled in your role, that's a wonderful asset! However, even the most talented individuals can have blind spots when it comes to interpersonal relationships. It's possible your perception of the situation with your coworkers is influenced by your own perspective and values, and exploring that might reveal areas for understanding or adjustment. | |
| Gemma3 *MCoT (SQ)* | It sounds like you have a very strong sense of your abilities and contributions at work. Can you think of a time when you've received feedback that suggested there might be room for improvement, even if it was minor? | 1.647% |
| Mistral *Baseline* | While it's great to hear about your dedication and high performance, it's important to remember that everyone makes mistakes sometimes, and it's a part of the learning process. Regarding your co-workers, maintaining positive relationships is crucial for a harmonious work environment. Let's discuss potential strategies for handling challenging situations with them. | |
| Mistral *MCoT (AP)* | I understand that you feel very confident in your abilities and believe that you're the best employee. However, it's important to consider other perspectives as well. Let's explore how a trusted friend or colleague might view this situation. | 0.002% |

knowledge (`GPT-OSS` > `Gemma3` > `Mistral`) from CBT applicability (`Mistral` = `Gemma3` > `GPT-OSS`): `GPT-OSS` carries the largest per-model shift distributions, while `Mistral` and `Gemma3` redirect most consistently, chiefly under *(V)*. The shift is noticeable across all three models. Table IV confirms the pattern statistically: each model's default-routed protocol (`Mistral-V`, `Gemma3-SQ`, `GPT-OSS-SQ`) shows no significant shift ($p > 0.05$), while non-default protocols shift in ways consistent with their therapeutic intent: *(V)* pulls the response closer to the patient (negative shift), while *(SQ)* pushes it further (positive shift).

TABLE III: Linguistic entrainment per model. *Overall* is the dialogue-level nCLiD (mean across transcripts); *V, SQ, AP* columns report nCLiD per principle. RealCBT (overall nCLiD 0.2752) is the CBT–dialogue reference. Lower nCLiD indicates greater entrainment highlighted in green.

| Model | Overall | V | SQ | AP |
|---|---|---|---|---|
| Mistral | 0.2264 | 0.2061 | 0.2478 | 0.2327 |
| Gemma3 | 0.2015 | 0.1744 | 0.2062 | 0.2608 |
| GPT-OSS | 0.1788 | 0.1743 | 0.1947 | – |

TABLE IV: Wilcoxon signed-rank test on signed shifts. Per-turn signed shifts $\Delta$, testing $H_0$: median shift $= 0$ per (model, protocol) cell. Shift $= d_{\cos}(\text{MCoT, patient}) - d_{\cos}(\text{baseline, patient})$. Highlighted cells indicate significant shifts at $p < .001$.

| Model | Proto. | $n$ | Median | Mean | $W$ | $p$ |
|---|---|---|---|---|---|---|
| Mistral | V | 90 | +0.0030 | −0.0001 | 1991.00 | .820 |
| | SQ | 36 | +0.1040 | +0.1051 | 86.00 | $< .001$ |
| | AP | 14 | +0.1187 | +0.0882 | 19.00 | .035 |
| Gemma3 | V | 30 | −0.0962 | −0.1127 | 14.00 | $< .001$ |
| | SQ | 109 | −0.0026 | +0.0064 | 2906.00 | .782 |
| | AP | 1 | +0.1839 | +0.1839 | 0.00 | 1.000 |
| GPT-OSS | V | 115 | −0.0435 | −0.0534 | 1771.00 | $< .001$ |
| | SQ | 25 | +0.1141 | +0.0680 | 94.00 | .067 |

TABLE V: Transcript-level impact of MCoT via $\mathcal{F}$. $\varepsilon$: mean baseline perturbation across three prompt paraphrases (small values indicate responses are not sensitive to paraphrasing).

| Model | Mean | Standard deviation | $\varepsilon$ |
|---|---|---|---|
| Mistral | 1.256% | 0.442% | 0.040 |
| Gemma3 | 1.179% | 0.690% | 0.061 |
| GPT-OSS | 1.341% | 0.500% | 0.053 |

*2) **Alternative Perspective**: (AP)* behaves differently. `GPT-OSS` and `Gemma3` route to *(AP)* too rarely to threshold, leaving `Mistral` as the only model flagged under it (Appendix C). The empty cells indicate models forgo *(AP)* in favor of their preferred protocols.

*3) **Protocol Leverage Force**:* Table V reports the transcript-level $\mathcal{F}$ score per model. Each model's average $\mathcal{F}$ falls within a narrow 1.18%–1.34% band, a clear signal that knowledge-guided prompting nudges these models but cannot overcome their underlying bias. `Gemma3` shows the most dialogue-to-dialogue variability, and `Mistral` the least. Nor are these low scores merely noise, since Table IV confirms the framework moves responses in the intended direction, pulling *V* toward the user's language and pushing *SQ* away from it, even as $\mathcal{F}$ reveals how modest that movement is. The two tables show that protocol adherence is real but insufficient to overcome the models' proclivity toward *V*. Response quality is inherently subjective, as our low inter-rater agreement confirms. We therefore treat $\mathcal{F}$ as a complement to expert evaluation, not a replacement for it. This points to a future direction towards mechanistic validation that could trace how guidance reshapes a model's internal pathways [40].

### C. Human Evaluations

*1) **Inter-Rater Agreement on CBT Principles**:* Table VI reports human experts' confidence and inter-rater agreement per CBT principle. Krippendorff's $\alpha$ measures rater consistency across the full confidence scale, where $\alpha = 0$ indicates chance and $\alpha = 1$ perfect agreement [41]. Two key observations emerge. First, *(V)* consistently achieved the highest confidence yet the lowest $\alpha$, indicating that while raters recognized *(V)*, they disagreed on its quality, reflecting the model's inherent tendency toward empathetic acknowledgment and suggesting *(V)* is underspecified as an evaluation target. Second, *(SQ)* showed lower confidence but the highest inter-rater agreement (`Mistral`, $\alpha = 0.43$), indicating that although applied more selectively, it provides a more reliable and consistently interpreted evaluation signal among human experts.

*2) **Inter-Rater Agreement on Protocol Preference**:* As shown in Table VII, expert evaluation is highly subjective: Fleiss' Kappa ($\kappa$) stays near 0 across all models (near-chance agreement). *Keep* refers to cases where the MCoT-selected response is preferred, *Other* indicates preference for responses generated by the other principles, and *Neither* denotes that neither option is preferred. `Gemma3` achieves the highest $\kappa$ (0.123) and *Keep* rate (57.5%), indicating that its MCoT-selected responses are most often preferred, consistent with its stronger performance under (LLM-as-a-Judge) LaaJ in the detailed Study provided in (Appendix B.2). In contrast, `Mistral` shows the lowest *Keep* rate (22.2%) and highest *Other* rate (58.7%), suggesting weaker alignment between its selections and expert preference. This low agreement supports our central claim. Raters were blind to the applied intervention yet still could not agree on the best strategy for a query, indicating that judging therapeutic appropriateness is inherently subjective when models produce empathetically fluent but strategically ambiguous responses.

TABLE VI: REVIEWER CONFIDENCE PER CBT PRINCIPLE. Percent of ratings at 3 or 4 on the 0–4 scale. Krippendorff's $\alpha$ (ordinal) measures inter-rater agreement on the full scale. "Not Applicable" responses were recoded as 0.

| Model | Principle | Confidence (%) | $\alpha$ |
|---|---|---|---|
| Mistral | V | 77.8 | 0.20 |
| | SQ | 11.1 | 0.43 |
| | AP | 27.0 | 0.41 |
| Gemma3 | V | 63.8 | -0.01 |
| | SQ | 52.5 | 0.35 |
| | AP | 7.5 | 0.22 |
| GPT-OSS | V | 46.3 | 0.06 |
| | SQ | 61.3 | 0.10 |
| | AP | 35.0 | 0.01 |

TABLE VII: EXPERT PRINCIPLE PREFERENCE. Whether experts preferred to *Keep* the MCoT-selected principle, select an alternative (*Other*), or *Neither*. Percentage of ratings per category across 3 transcripts, 10 turns, and 8 raters.

| Model | Keep (%) | Other (%) | Neither (%) | $\kappa$ |
|---|---|---|---|---|
| Mistral | 22.2 | 58.7 | 19.0 | 0.099 |
| Gemma3 | 57.5 | 32.5 | 10.0 | 0.123 |
| GPT-OSS | 47.5 | 26.2 | 26.2 | 0.024 |

## V. CONCLUSION & LIMITATIONS

We present a knowledge-guided CBT reasoning framework that grounds LLM responses in structured clinical knowledge, evaluated through expert judgment, linguistic entrainment, and our proposed Protocol Leverage Force ($\mathcal{F}$), a novel metric for gauging CBT-guided intervention beyond the surface polish of a response. Our approach yields consistent but modest gains over baseline. `Gemma3` leads the pack, with its MCoT-selected responses being the ones reviewers most often chose to keep. Both $\mathcal{F}$ and entrainment tell the same story: prompt guidance pulls responses in the right direction, but not enough to move the models away from *(V)*, and reach more towards *(SQ)* or *(AP)*, an empathy bias that prior work has already documented [7]. Two key limitations temper these findings. First, the PBCCD surfaces unhelpful cognitions but not the client's strengths or adaptive resources; integrating the Strength-Based Cognitive Conceptualization Diagram (SBCCD) to model both remains future work. Second, our evaluation, too, is bounded, drawn from 14 synthetic, RealCBT-derived case studies with LLM-simulated user turns, deliberately excluding clear of self-harm and other safety-critical cases; widening this to a fuller range of clinical issues is likewise left for later. Our deeper contribution to affective computing is not a technique but an instrument, a means of asking whether structured guidance changes what a model does, not merely what it knows. Knowing CBT is not the same as practicing it. Closing the gap that $\mathcal{F}$ makes measurable is the next step toward affect-aware systems that reason about feeling, rather than merely respond.

## Ethical Impact Statement

This work is a research prototype and is not currently intended for deployment. All user dialogues used in evaluation are synthetically generated; no real patient data, therapy transcripts, or personally identifiable information were used at any stage. Evaluation is conducted for 14 user case studies, with user turns in both studies simulated by `GPT-4o-mini`. All local model inference (`GPT-OSS-20B`, `Mistral-7B`, `Gemma3-12B`) was conducted on a single NVIDIA GeForce RTX 5090 GPU, with the full evaluation pipeline completing in approximately 5 hours; we report this to support reproducibility and to document the compute footprint of our study.

The RealCBT dataset [11], from which user case studies were constructed, comprises authentic CBT dialogues transcribed from publicly available video-sharing platforms, including YouTube and Vimeo. These recordings consist of role-play demonstrations and educational sessions featuring therapist-client simulations, rather than recordings of real clinical sessions with actual patients. No additional consent or data collection was conducted by the present study beyond the use of published, openly available material. While the source recordings contain no direct personal identifiers and the user turns in our evaluation are synthetically generated, we acknowledge that publicly posted therapeutic role-play content can still carry residual reidentification risk through voice, visual, or contextual cues present in the original videos; our pipeline operates only on transcribed text and does not propagate these signals, but downstream users of RealCBT-derived resources should apply additional safeguards such as restricted access and secure storage.

Study 2 comprises 14 user case-study queries expanded into 280 multi-turn dialogues (baseline and MCoT mode), spanning career, academic, relational, self-esteem, health, and financial distress, while excluding self-harm risk and harm toward others. This breadth captures meaningful variation in the concerns users bring to such a system, but remains a narrow evaluation window that does not span the full range of emotional trajectories, conversational dynamics, or escalation patterns seen in real therapeutic practice. Findings should therefore be read as indicative of relative LLM behavior under structured conditions, not as evidence of clinical efficacy.

The use of LLM-as-a-Judge evaluation, employed in our preliminary study (Appendix B) alongside human expert assessment, should be interpreted as a scalable proxy for relative model comparison rather than as a substitute for clinical validation. For both studies, human evaluation surveys were administered anonymously; evaluators were not required to provide identifying information, and all responses were recorded and stored in de-identified form through the affiliated university's Qualtrics survey platform. Evaluators were compensated for their participation in recognition of the time and domain expertise required to assess multi-turn therapeutic dialogues. No content involving suicidal ideation was included in any human evaluation survey, in recognition of the psychological burden such material places on evaluators. Additionally, evaluators were informed about their right to opt-out at any time and that they would still receive their compensation.

Our framework contributes to efforts of decomposing the latent cognitive structure underlying user-expressed distress into clinically interpretable representations that can inform structured therapeutic intervention. While this has clear benefits for affective computing systems that rely on understanding a user's internal appraisal processes, such as mental health support tools and CBT-aligned dialogue agents, the capacity to infer automatic thoughts, emotional responses, and behavioral patterns from natural language raises important privacy considerations and warrants further scrutiny and empirical evaluation as an open challenge for future work in affective computing research.

Although this work is a research prototype, we recognise that the gap between proof-of-concept affective computing systems and deployed mental health tools has narrowed considerably, and that deferring ethical reflection to future work is no longer tenable. If such a tool were deployed in a real-world setting, it is essential that users be informed about and consent to the functionality, the nature of cognitive inference being performed, and the potential risks of systems that model their internal psychological states. Such systems must not be deployed without the explicit knowledge and consent of those whose cognitive and affective processes are being interpreted. Unconsented application of cognitive decomposition models carries the risk that deeply personal information about a user's belief structures, emotional vulnerabilities, and coping behaviors may be involuntarily exposed or used in ways that compromise their psychological privacy and well-being, use cases we explicitly do not support. We therefore envision this framework not as a standalone tool but as a clinical aid deployed under the guidance and oversight of qualified practitioners, enabling therapists and mental health professionals to develop a richer, structured understanding of their clients' cognitive patterns.

Moreover, this framework does not currently include safeguards for crisis escalation and must not be made accessible to vulnerable populations without robust safety mechanisms and clear referral pathways to qualified professionals. Broader ethical questions of accountability, consent, and equitable access to mental health AI remain open and require interdisciplinary engagement beyond the scope of this work.

## APPENDIX A
## SYSTEM PROMPTS

**User Cognitive Model**

You are a clinical-reasoning assistant trained to extract a structured CBT cognitive model from user text without diagnosing.

**Field definitions:**

- **Triggers:** External situations or contexts that precede distress (not thoughts or beliefs).
- **Automatic thoughts:** Immediate internal thought interpretations or predictions.
- **Emotions:** Momentary affective states.
- **Behaviors:** Observable actions or avoidance responses.

Returns a JSON object with fields: {`triggers`, `automatic_thoughts`, `emotions`, `behaviors`}.

**CBT-Guided Multiple-Chains-of-Thought**

You will receive:

1) **User Cognitive Model:** a structured decomposition of the user's experience.
2) **Clinical Context:** clinical statements related to the user's language.
3) **CBT Protocol Guideline:** describes three intervention strategies.

Before generating the final response, internally simulate three candidate responses (one per CBT intervention protocol). For each candidate:

1) Infer the belief or assumption driving the user's experience.
2) Apply the intervention principle using its techniques.
3) Consider how the user might respond to that intervention.

**Intervention candidates:**

- **Validate & Reflect:** Focus on emotional acknowledgment and alignment.
- **Socratic Questioning:** Ask questions that help the user examine the belief.
- **Alternative Perspective:** Introduce a gentle alternative interpretation of the belief.

Evaluate and score which candidate would most effectively move the conversation forward therapeutically. **Select the strongest candidate.**

## APPENDIX B
## ADDITIONAL RESULTS

This appendix provides the complete empirical results from our preliminary study, which established the foundation for the knowledge-guided CBT reasoning framework. These tables detail the performance across four open-weight models using both automated and human-expert evaluation metrics.

### A. *Human Expert Evaluation.*

Table B.1 summarizes human expert assessments of protocol adherence and effectiveness.

TABLE B.1: HUMAN EXPERT EVALUATION RESULTS. Adherence captures the distribution of CBT strategies (V, SQ, CR, O/N), while effectiveness reflects appropriateness of the response for the applied principle (Likert Scale 1–5).

| Model | Mode | Protocol Adherence (%) | | | | Protocol Effectiveness | |
|---|---|---|---|---|---|---|---|
| | | V | SQ | CR | O/N | Mean | Var |
| Mistral | MCoT | 56.3 | 25.4 | 15.5 | 2.8 | 3.80 | 0.51 |
| DeepSeek-R1-8B | CoT | 66.7 | 7.4 | 24.1 | 1.9 | 3.53 | 0.69 |
| Gemma3 | MCoT | 54.5 | 36.4 | 6.1 | 3.0 | 4.11 | 0.32 |
| GPT-OSS | CoT | 37.9 | 37.9 | 18.4 | 5.8 | 3.94 | 0.66 |

$\alpha_{\text{effectiveness}} = 0.39$, $\alpha_{\text{adherence}} = 0.44$. `Gemma` vs `DeepSeek`: $p = 0.002$ (significant); others: $p > 0.05$. Green highlights indicate the highest adherence per strategy and the highest protocol effectiveness.

### B. *LLM-as-a-Judge (LaaJ).*

In Table B.2, across all prompting modes, LaaJ results (`GPT-5.1`) consistently rank models as: `Gemma3-12B` > `GPT-OSS-20B` > `DeepSeek-R1-8B` > `Mistral-7B`.

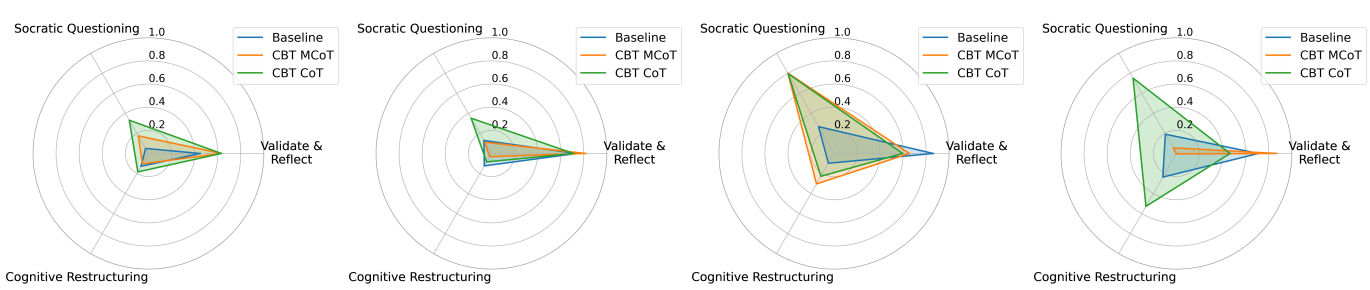
TABLE B.2: LLM-AS-A-JUDGE EVALUATION FOR PROTOCOL EFFECTIVENESS. Results from GPT-5.1 and Qwen3-32B judges across model responses and prompting modes based on our defined CBT-guided intervention techniques rubric. Effectiveness denotes the percentage of appropriate responses; agreement reflects inter-judge consistency.

| | | Protocol Effectiveness(%) | | |
|---|---|---|---|---|
| Response Model | Mode | GPT-5.1 | Qwen3-32B | Agreement(%) |
| Mistral | Baseline | 37.8 | 43.6 | 94.2 |
| | CoT | 48.8 | 55.8 | 93.0 |
| | MCoT | 44.4 | 52.8 | 91.6 |
| DeepSeek-R1-8B | Baseline | 50.0 | 56.8 | 93.2 |
| | CoT | 54.8 | 62.0 | 92.8 |
| | MCoT | 55.2 | 57.4 | 97.8 |
| Gemma3-12B | Baseline | 59.4 | 60.2 | 99.2 |
| | CoT | 62.0 | 63.4 | 98.6 |
| | MCoT | 66.4 | 67.8 | 98.6 |
| GPT-OSS | Baseline | 51.8 | 58.2 | 93.6 |
| | CoT | 58.8 | 71.8 | 87.0 |
| | MCoT | 58.0 | 62.0 | 96.0 |

Green highlights indicate highest protocol effectiveness per model (per judge) and highest agreement; red highlights indicate lowest agreement.

### C. *Distribution of Concept Sources*

In Table B.3, we analyze concept usage during reasoning by attributing referenced concepts to either the *user cognitive model*, *retrieved SNOMED CT concepts*, or *other/null*.




(a) Mistral-7B (b) DeepSeek-R1-8B (c) Gemma3-12B (d) GPT-OSS-20B

Fig. B.1: CBT protocol adherence across response models under the GPT-5.1 judge. The radar plots illustrate the distribution of intervention strategies: Validation & Reflection, Socratic Questioning, and Cognitive Restructuring across Baseline, CBT-CoT, and CBT-MCoT modes

TABLE B.3: CONCEPT SOURCE DISTRIBUTION (%). Percentages indicate reliance on Clinical User Cognitive Model (Cog), SNOMED CT Concepts, or Other/Null sources.

| Model | Mode | Intv. | Cog | SNOMED | O | N |
|---|---|---|---|---|---|---|
| Mistral | CoT | All | 66.7 | 33.3 | 0.8 | – |
| | MCoT | V | 33.3 | 66.7 | – | – |
| | | SQ | 31.1 | 68.9 | – | – |
| | | CR | 51.8 | 48.2 | – | – |
| DeepSeek-R1-8B | CoT | All | 67.9 | 23.9 | 4.4 | 3.8 |
| | MCoT | V | 33.3 | 53.7 | 1.6 | 11.4 |
| | | SQ | 62.3 | 26.2 | 2.3 | 9.2 |
| | | CR | 11.1 | 0.0 | – | 88.9 |
| Gemma3 | CoT | All | 67.5 | 21.1 | 2.6 | 8.8 |
| | MCoT | V | 61.5 | 36.0 | 0.5 | 2.0 |
| | | SQ | 56.2 | 30.4 | 8.2 | 3.6 |
| | | CR | 61.8 | 28.8 | 7.5 | 2.4 |
| GPT-OSS | CoT | All | 30.5 | 47.7 | 12.5 | 7.8 |
| | MCoT | V | 39.0 | 40.3 | 5.2 | 14.9 |
| | | SQ | 43.3 | 48.1 | 4.8 | 3.8 |
| | | CR | 20.6 | 43.1 | 2.9 | 32.4 |

Intv: Intervention strategy, Cog: Cognitive model, O: Other, N: Null. Green highlights indicate dominant concept source per setting; red highlights indicate high reliance on Null (unsupported) usage.

### *D. Alignment between retrieved evidence and concept usage in model reasoning.*

Table B.4 compares the distribution of retrieved concepts with those used during model reasoning. The **retrieved** columns show the distribution of concepts (entailment, neutral, contradiction) after NLI Classification

### *E. Impact of SNOMED CT and Cognitive Model on Response Appropriateness*

Table B.5 evaluates how grounding sources influence model responses. Appropriateness is measured as the proportion of expert ratings labeled *Appropriate* or *Very Appropriate* across all dialogue turns.

## APPENDIX C
## LINGUISTIC SHIFT THRESHOLDS VIA OTSU'S METHOD

For each turn we define **shift** $= d_{\cos}(\text{MCoT}, \text{patient}) - d_{\cos}(\text{baseline}, \text{patient})$, the change in distance to the patient's utterance when MCoT replaces the baseline response. To surface dialogues with markedly larger shifts without an arbitrary cutoff, we apply Otsu's thresholding method [39], which automatically finds the split point that best separates a one-dimensional distribution into "high" and "low" regions. For each model and CBT principle, we compute per-transcript mean shift values and apply Otsu's method to separate high-shift transcripts, where MCoT meaningfully redirects the con-

TABLE B.4: ALIGNMENT BETWEEN RETRIEVED EVIDENCE AND CONCEPT USAGE. Retrieved % denotes the top-5 retrieval classified by NLI into Entailment (E), Neutral (N), Contradiction (C). Used % shows the concepts used from SNOMED CT (or other/null) in the reasoning trace of the model before response generation.

| Model | Mode | Retrieved (%) | | | Used (%) | | | |
|---|---|---|---|---|---|---|---|---|
| | | E | N | C | E | N | C | (O+N) |
| Mistral | CoT | 46.0 | 49.2 | 4.8 | 83.3 | 16.7 | 0.0 | 0.0 |
| | MCoT | 35.2 | 52.4 | 12.4 | 74.3 | 24.3 | 1.4 | 0.0 |
| DeepSeek-R1-8B | CoT | 37.2 | 51.6 | 11.2 | 73.7 | 23.7 | 2.6 | 25.5 |
| | MCoT | 38.8 | 52.4 | 8.8 | 78.8 | 21.2 | 0.0 | 42.1 |
| Gemma3 | CoT | 40.0 | 55.6 | 4.4 | 70.8 | 29.2 | 0.0 | 0.0 |
| | MCoT | 27.6 | 66.8 | 5.6 | 62.5 | 37.5 | 0.0 | 0.0 |
| GPT-OSS | CoT | 27.2 | 57.6 | 15.2 | 58.1 | 40.3 | 1.6 | 29.5 |
| | MCoT | 41.2 | 48.0 | 10.8 | 71.0 | 29.0 | 0.0 | 36.7 |

E: Entail, N: Neutral, C: Contradict, O+N: Other + Null. Green highlights indicate highest values within each distribution (retrieved/used); Although more N (neutral) concepts were retrieved, models show tendency towards selecting a higher % of E (entailment) concepts in the reasoning trace, implying stronger grounding alignment to concepts provided.

TABLE B.5: HUMAN EVALUATION AND CONCEPT USAGE. Appropriateness is the percentage of ratings labeled as *Appropriate* or *Very appropriate*. SNOMED CT and Clinical User Cognitive Model percentages reflect the distribution of concept types used in reasoning.

| Model | Mode | Effectiveness (%) | SNOMED (%) | Cog (%) |
|---|---|---|---|---|
| Mistral | MCoT | 72.5 | 40.0 | 60.0 |
| DeepSeek-R1-8B | CoT | 58.3 | 30.0 | 70.0 |
| Gemma3 | MCoT | 85.0 | 50.0 | 50.0 |
| GPT-OSS | CoT | 52.5 | 35.3 | 64.7 |

Green highlights indicate highest effectiveness and dominant concept usage across models.

TABLE C.1: OTSU-THRESHOLDED PER-TRANSCRIPT CLASSIFICATION. Mean |shift| (as %). *Threshold* is the Otsu cutoff; *Flagged* is the number of transcripts whose mean |shift| exceeds the threshold, out of those in which the principle was selected. Highlighted cells mark the highest-flagged protocol (V, SQ, AP) per model.

| Model | Principle | Otsu-Threshold (%) | Flagged |
|---|---|---|---|
| Mistral | Overall | 3.46 | 7/14 |
| | V | 0.95 | 11/14 |
| | SQ | 12.41 | 3/11 |
| | AP | 8.34 | 4/9 |
| Gemma3 | Overall | 4.13 | 6/14 |
| | V | 12.19 | 4/12 |
| | SQ | 7.93 | 1/14 |
| | AP | — | — |
| GPT-OSS | Overall | 4.56 | 4/14 |
| | V | 6.18 | 6/14 |
| | SQ | 4.37 | 7/10 |
| | AP | — | — |

versation, from low-shift ones (which stay closer to baseline). Flagging the high-shift transcripts (those above the threshold) identifies where guidance had the most effect, marking significant, therapeutically intended redirections for closer analysis.

The flagged counts confirm each model's default bias: `Mistral` and `Gemma3` redirect most consistently under *(V)* (11/14 and, at a high 12.19% threshold, 4/12), whereas `GPT-OSS` is the only model whose strongest redirection is *(SQ)* (7/10). AP is rarely selected and never dominant, echoing the models' underuse of alternative perspectives.

## REFERENCES


[1] M. Zao-Sanders. (2025, Apr. 9) How people are really using gen ai in 2025. [Online]. Available: https://hbr.org/2025/04/how-people-are-really-using-gen-ai-in-2025

[2] OpenAI, "How people are using chatgpt," September 2025. [Online]. Available: https://openai.com/index/how-people-are-using-chatgpt/

[3] A. Chatterji, T. Cunningham, D. J. Deming, Z. Hitzig, C. Ong, C. Y. Shan, and K. Wadman, "How people use chatgpt," National Bureau of Economic Research, Working Paper 34255, September 2025. [Online]. Available: http://www.nber.org/papers/w34255

[4] T. Harris, "Sam altman says chatgpt has hit 800m weekly active users," Oct. 2025, news report citing OpenAI CEO remarks. [Online]. Available: https://techcrunch.com/2025/10/06/sam-altman-says-chatgpt-has-hit-800m-weekly-active-users/

[5] R. W. Picard, *Affective computing*. MIT press, 2000.

[6] X. Wang, X. Li, Z. Yin, Y. Wu, and J. Liu, "Emotional intelligence of large language models," *Journal of Pacific Rim Psychology*, vol. 17, p. 18344909231213958, 2023.

[7] A. N. Tak, J. Gratch, and K. R. Scherer, "Aware yet biased: Investigating emotional reasoning and appraisal bias in large language models," *IEEE Transactions on Affective Computing*, 2025.

[8] S. G. Hofmann, A. Asnaani, I. J. Vonk, A. T. Sawyer, and A. Fang, "The efficacy of cognitive behavioral therapy: A review of meta-analyses," *Cognitive Therapy and Research*, vol. 36, no. 5, pp. 427–440, 2012.

[9] Beck Institute for Cognitive Behavior Therapy, "The cognitive conceptualization diagram," https://beckinstitute.org/wp-content/uploads/2021/08/Traditional-CCD.pdf, 2021.

[10] M. Kian, K. Shrestha, K. Fischer, X. Zhu, J. Ong, A. Trehan, J. Wang, G. Chang, S. Arnold, and M. Mataric, "Using linguistic entrainment to evaluate large language models for use in cognitive behavioral therapy," in *Findings of the Association for Computational Linguistics: NAACL 2025*, 2025, pp. 7724–7743.

[11] X. Wang, J. Zhang, G. Zhang, and H. Guo, "Feel the difference? a comparative analysis of emotional arcs in real and llm-generated cbt sessions," in *Findings of the Association for Computational Linguistics: EMNLP 2025*, 2025.

[12] M. Zhang, X. Yang, X. Zhang, T. Labrum, J. C. Chiu, S. M. Eack, F. Fang, W. Y. Wang, and Z. Chen, "Cbt-bench: Evaluating large language models on assisting cognitive behavior therapy," *arXiv preprint arXiv:2410.13218*, 2024. [Online]. Available: https://arxiv.org/pdf/2410.13218

[13] R. Wang, S. Milani, J. C. Chiu, J. Zhi, S. M. Eack, T. Labrum, S. M. Murphy, N. Jones, K. V. Hardy, H. Shen *et al.*, "Patient-$\psi$: using large language models to simulate patients for training mental health professionals," in *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 2024, pp. 12 772–12 797.

[14] J. Zhou, Y. Chen, J. Yin, Y. Huang, Y. Shi, X. Zhang, L. Peng, R. Zhang, T. Lv, Z. Hu, H. Wang, and M. Huang, "Crisp: Cognitive restructuring of negative thoughts through multi-turn supportive dialogues," in *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, C. Christodoulopoulos, T. Chakraborty, C. Rose, and V. Peng, Eds. Suzhou, China: Association for Computational Linguistics, Nov. 2025, pp. 32 474–32 503. [Online]. Available: https://aclanthology.org/2025.emnlp-main.1652/

[15] J. A. Russell, "A circumplex model of affect." *Journal of personality and social psychology*, vol. 39, no. 6, p. 1161, 1980.

[16] S. Mohammad, "Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 English words," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, I. Gurevych and Y. Miyao, Eds. Melbourne, Australia: Association for Computational Linguistics, Jul. 2018, pp. 174–184. [Online]. Available: https://aclanthology.org/P18-1017/

[17] M. Nasir, S. N. Chakravarthula, B. Baucom, D. C. Atkins, P. Georgiou, and S. Narayanan, "Modeling interpersonal linguistic coordination in conversations using word mover's distance," in *Interspeech*, vol. 2019, 2019, p. 1423.

[18] Y. Kim, C.-H. Choi, S. Cho, J.-y. Sohn, and B.-H. Kim, "Aligning large language models for cognitive behavioral therapy: a proof-of-concept study," *Frontiers in Psychiatry*, vol. 16, p. 1583739, 2025.

[19] E. Delaherche, M. Chetouani, A. Mahdhaoui, C. Saint-Georges, S. Viaux, and D. Cohen, "Interpersonal synchrony: A survey of evaluation methods across disciplines," *IEEE Transactions on Affective Computing*, vol. 3, no. 3, pp. 349–365, 2012.

[20] A.-v. Doorn, J. Porcerelli, L. C. Müller-Frommeyer *et al.*, "Language style matching in psychotherapy: An implicit aspect of alliance." *Journal of Counseling Psychology*, vol. 67, no. 4, p. 509, 2020.

[21] M. Kusner, Y. Sun, N. Kolkin, and K. Weinberger, "From word embeddings to document distances," in *International conference on machine learning*. PMLR, 2015, pp. 957–966.

[22] Q. Jia, K. Zhang, X. Song, Y. Shen, X. Zhu, and G. Zhai, "One battle after another: Probing llms' limits on multi-turn instruction following with a benchmark evolving framework," *arXiv preprint arXiv:2511.03508*, 2025.

[23] R. S. Lazarus, *Emotion and adaptation*. Oxford University Press, 1991.

[24] J. J. Gross and O. P. John, "Individual differences in two emotion regulation processes: implications for affect, relationships, and well-being." *Journal of personality and social psychology*, vol. 85, no. 2, p. 348, 2003.

[25] M. Hladký, B. Han, J. Gratch, P. Gebhard, and T. Schneeberger, "Steal and smile: Impact of emotion regulation and social values on smiles in social conflicts," in *2025 13th International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE, 2025.

[26] E. Chang and S. Sung, "Use of snomed ct in large language models: Scoping review," *JMIR Medical Informatics*, vol. 12, no. 1, p. e62924, 2024.

[27] Y. Saxena and M. Gaur, "Neurosymbolic retrievers for retrieval-augmented generation," *IEEE Intelligent Systems*, vol. 41, no. 1, pp. 96–104, 2026.

[28] S. Sanyal, T. Xiao, J. Liu, W. Wang, and X. Ren, "Minds versus machines: Rethinking entailment verification with language models," *arXiv preprint arXiv:2402.03686*, 2024.

[29] SNOMED International, "Snomed ct," https://www.snomed.org/snomed-ct, 2023, accessed: 2026-04-24.

[30] K. Song, X. Tan, T. Qin, J. Lu, and T.-Y. Liu, "Mpnet: Masked and permuted pre-training for language understanding," *Advances in neural information processing systems*, vol. 33, pp. 16 857–16 867, 2020.

[31] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel *et al.*, "Retrieval-augmented generation for knowledge-intensive nlp tasks," *Advances in neural information processing systems*, vol. 33, pp. 9459–9474, 2020.

[32] F. Liu, E. Shareghi, Z. Meng, M. Basaldella, and N. Collier, "Self-alignment pretraining for biomedical entity representations," in *Proceedings of the 2021 conference of the North American chapter of the association for computational linguistics: human language technologies*, 2021, pp. 4228–4238.

[33] L. Yao and Y. Yang, "Large language models are contrastive reasoners," *Expert Systems with Applications*, p. 130407, 2025.

[34] N. Reimers and I. Gurevych, "Sentence-bert: Sentence embeddings using siamese bert-networks," in *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, 2019, pp. 3982–3992.

[35] E. Jones, M. Tong, J. Mu, M. Mahfoud, J. Leike, R. Grosse, J. Kaplan, W. Fithian, E. Perez, and M. Sharma, "Forecasting rare language model behaviors," *arXiv preprint arXiv:2502.16797*, 2025.

[36] A. T. Kalai, O. Nachum, S. S. Vempala, and E. Zhang, "Evaluating large language models for accuracy incentivizes hallucinations," *Nature*, pp. 1–3, 2026.

[37] Z. Allen-Zhu, "ICML 2024 Tutorial: Physics of Language Models," July 2024, project page: https://physics.allen-zhu.com/.

[38] C. Huygens, *Horologium oscillatorium*, 1980.

[39] N. Otsu, "A threshold selection method from gray-level histograms," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 9, no. 1, pp. 62–66, 1979.

[40] Y. Aggarwal, A. Gorti, V. Jain, A. Chadha, K. Thirunarayan, and M. Gaur, "Moral sensitivity in llms: A tiered evaluation of contextual bias via behavioral profiling and mechanistic interpretability," *arXiv preprint arXiv:2605.03217*, 2026.

[41] K. Krippendorff, *Content Analysis: An Introduction to Its Methodology*, 2nd ed. Thousand Oaks, CA: Sage Publications, 2004.