Cognitive Distortion Detection with LLM-generated Datasets

Anonymous ACL submission

Abstract

We present a novel framework for simulating and detecting cognitive distortions (CoDs) in therapist–patient dialogues using large language models (LLMs) and structured therapeutic simulations. By creating individualized distortion profiles for patients and prompting LLMs based on Cognitive Behavioral Therapy (CBT) principles, we simulate therapy sessions that undergo iterative refinement in a reward-guided loop, maximizing naturalness, coherence, and alignment with targeted distortions. We then introduce inline CoD annotations as weak supervision and assess their effect on classifier performance. Leveraging both LLM-simulated sessions and a public CoD dataset through hybrid embeddings, our approach achieves a 0.74 weighted F1. These findings highlight the promise of controlled simulation and iterative reinforcement to boost data-scarce clinical NLP tasks.

1 Introduction

Cognitive distortions (CoDs) are central to mental health analysis and therapy. Their contextual analysis supports effective intervention; for example, Na (2024) used LLMs with CBT-based prompts to build a QA dataset. Prior work has identified CoDs in clinical dialogue (Shreevastava and Foltz, 2021; Singh et al., 2024) and explored LLMs as counselors for their conversational fluency (Raile, 2024; Berrezueta-Guzman et al., 2024; Pirnay, 2023-04-27). However, many such systems transmit sensitive user data to third-party providers.

We address the need for secure, scalable CoD data by introducing a method for simulating realistic, labeled therapy sessions using API-access LLMs (Figure 1). Our approach simulates individualized CoD profiles, going beyond prior work on emotional rapport or symptoms (Wang et al., 2024), and tailors therapy sessions toward distorted cognition. The result is a synthetic dataset of therapist–patient dialogues, with patient responses guided by weighted distortion profiles reflecting realistic severity levels. Our contributions include:

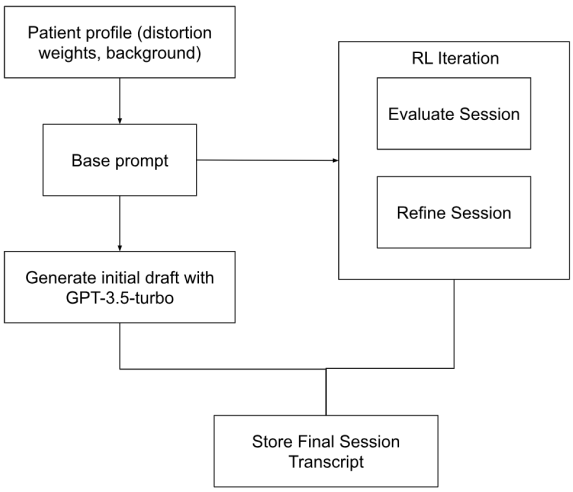

Figure 1: Reward guided Reinforcement Learning loop for LLM generated transcripts.

- We propose an approach for generating realistic synthetic data to support CoD detection using individualized distortion profiles and cognitive behavioral therapy principles.
- We introduce inline CoD annotations as a weak supervision method for CoD detection.
- We contribute a new dataset of simulated therapist-patient dialogues, with patient responses guided by realistic, weighted distortion profiles of varying severity.

Our method instructs the LLM to act as a CBT-trained therapist, consistently applying structured interventions (Figure 2). We selected CBT due to its empirical grounding and compatibility with computational modeling. Unlike more interpretive or psychodynamic therapies, CBT focuses on specific, labelable thought patterns—ideal for training models to detect and refine distorted thinking.

Therapist	I'm glad to see you back for our second session. How have things been since our last meeting?
Patient	Well, it's been a rollercoaster, to be honest. I feel like [overgeneralization] everything I write is just terrible, and nobody will ever appreciate it.
Therapist	It sounds like you've been experiencing some overgeneralization in your thoughts about your writing. Let's try to unpack that a bit. Can you think of any instances where someone did appreciate your work, even if it's just a small moment?

Figure 2: Excerpt from a generated session with therapist intervention. Full tagged transcript in Appendix B.

2 Background

Cognitive Distortion data can boost mental disorder detection, enabling automated tools like annotation platforms and real-time therapy assistants. Despite their clinical importance, CoDs are under-explored in NLP due to limited and inconsistent datasets (Wang et al., 2023; Lybarger et al., 2022). We follow Shreevastava and Foltz (2021), focusing on the nine most salient CoDs (e.g., *overgeneralization* or *mind reading*) while omitting *magnification*, given its overlap with distortions like *catastrophizing*. Appendix A lists definitions and examples.

Our work extends prior studies (Chen et al., 2023) by systematically evaluating multiple classifiers trained on both public and LLM-generated data. It also introduces a reward-guided loop (Figure 1) that refines generated therapy dialogues for coherence, realism, and distortion relevance. Moreover, we embed inline CoD tags (e.g., `[overgeneralization]`) into patient utterances and assess their value for classification, exemplified in Figure 2.

Doctor-patient conversations were framed as QA tasks to link responses with clinical records (Arana et al., 2024). Agent-assisted dialogues were explored by Liu et al. (2024), showing how empathetic, context-aware exchanges boost patient engagement and outcomes. CBT's structured techniques and empirical grounding make it ideal for systematic CoD modeling in NLP. Prior work supporting the integration of LLMs with CBT principles for NLP applications include that of Wang et al. (2024); Lim et al. (2024); de Toledo Rodriguez et al. (2021), and Maddela et al. (2023), among others.

Listing 1: Sample Patient Profile

```
{
    "patient_id": 1,
    "name": "Patient1",
    "age": 37,
    "primary_cod": {
        "overgeneralization": 1,
        "mind reading": 9,
        "personalization": 6,
        "catastrophizing": 1,
        "all-or-nothing thinking": 2,
        "mental filter": 7,
        "disqualifying the positive": 9,
        "emotional reasoning": 9,
        "should statements": 3
    },
    "background": "Patient1 is a 37
        -year-old doctor who
        experiences mind reading due to
        the demands of their
        profession."
}
```

3 Dataset Structure

3.1 Virtual Patient Profiles

We created 100 simulated patient profiles using a custom prompt. Each profile has a unique ID, an age (18–80), a brief background, and a CoD profile with weighted scores (0–10) for nine distortions. Race, gender, and cultural details were omitted to avoid stereotypes, and only 33% have non-zero distortion weights for realistic prevalence. Listing 1 provides an example of a simulated patient profile. Appendix C details profile attributes, and Appendix D (Figure 5) shows the full prompt.

3.2 Cognitive Distortion Keyword Mapping

We categorized well-known distortions by type and then systematically mapped them to specific keyword patterns. Appendix A summarizes the distortion types. We used a transformer-based language model to simulate reward-guided multi-turn therapy sessions between each simulated patient and a simulated senior CoD therapist trained in CBT techniques, as described later.

3.3 Session Structure

Each simulated patient participated in up to three therapy sessions, where the therapist guided discussions based on patient profiles. A primary role is assigned to the LLM, instructing it to simulate a therapy session between a CBT-trained senior therapist and patient (see Appendix D, Figure 7). Session count is limited to a predefined number

Algorithm 1 Reward Calculation

Require: Session Text T
while $t \in T$ **do**
  $N \leftarrow \text{Naturalism}(T)$
  $C \leftarrow \text{Coherence}(T)$
  $D \leftarrow \text{DistortionMatch}(T)$
  $R \leftarrow \alpha N + \beta C + \gamma D$
end while
return $\{N, C, D, reward = R\}$

Algorithm 2 Generate Initial Transcript

Require: Developer prompt $role$
Require: user prompt $base\_prompt$
Ensure: Initial LLM response $draft\_output$
try:
  $response \leftarrow gpt(messages)$
  $draft\_output \leftarrow response$
except Exception e:
  Log error with patient name and session number **return** $response$, $draft\_output$

and we use the distortion dictionary to generate distortion-weighted patient responses.

4 Methodology

4.1 Distortion Injection and Control

We employed a targeted injection process to ensure diverse CoDs in each conversation, guided by each patient's unique distortion profile. Prompts instructed the LLM to naturally manifest the specified distortions (Figure 7), while a keyword-based mapping system (e.g., "nothing will ever work out" $\rightarrow$ catastrophizing) served as diagnostic targets during evaluation. A reward function checked alignment between the generated text and the distortion profile; transcripts failing to exhibit realistic distortions were refined until improvement was observed.

4.2 Simulating a Multi-Session Therapy Protocol with Rewards

We developed a multi-session therapy simulation grounded in CBT, leveraging patient profiles to test reinforcement-guided conversations. A reward-based loop computes scores for each session. Because we employ few-shot prompts to simulate multi-turn therapy, each prompt includes the patient's background, CoD profile, and previous-session notes (Appendix D, Figure 8).

As shown in Algorithm 1, each iteration evaluates the LLM-generated session based on a composite reward score R, which is composed of three weighted components: naturalism, coherence, and distortion match. If the new output yielded a reward higher than the current best ($R_{t+1} > R_t$), the session was retained for further refinement. Otherwise, the loop exits early to prevent degradation of session quality. In some implementations, refinement could continue until a 30% gain was observed, but we chose to break when no improvement was detected to favor efficiency and avoid overfitting.

The reward is defined as:

$$R^{(i)} = \alpha \cdot N^{(i)}(s) + \beta \cdot C^{(i)}(s) + \gamma \cdot D^{(i)}(s, P) \quad (1)$$

where $R^{(i)}$ is the total reward at iteration i, $N^{(i)}(s)$ is the naturalism score for session text s, $C^{(i)}(s)$ is the coherence score for session text s, $D^{(i)}(s, P)$ is the distortion alignment score given session s and patient profile P, and α, β, γ are scalar weights assigned to each component. In our default setup, we assign equal weights: $\alpha = \beta = \gamma = 1.0$, to reflect equal importance of fluency, structural consistency, and task relevance. These weights were chosen empirically based on pilot runs that showed no metric dominated quality across all examples.

During generation, a baseline reward R_0 is computed for the initial session. The refinement loop iterates to generate alternative sessions $S^{(i)}$, each with a corresponding reward $R^{(i)}_{\text{session}}$. The loop terminates if:

$$R^{(i)}_{\text{session}} \geq (1 + \delta) \cdot R_0 \quad (2)$$

where δ is the predefined improvement threshold (e.g., $\delta = 0.3$ for 30% improvement).

Once the best session is chosen, a progress note is generated summarizing its distortion profile: if CoDs appear, they're listed; otherwise, the note indicates no distortions. These notes serve as session summaries and maintain continuity. We calculate the total reward for a multi-turn session S as a weighted sum of naturalism, coherence, and distortion match, each measured over the entire session.

We simulate multi-session therapy by first generating an initial transcript with a system-defined role and base prompt (Algorithm 2). Each patient has weighted CoD predispositions (0.0–1.0) to shape their responses. A scoring rubric and fixed output format minimize LLM variance and enable consistent evaluation. These scores feed into the reward function, guiding the LLM to produce realistic, coherent therapy dialogues. For each session, the system defines a system role and a generation prompt incorporating the patient's background, previous notes, and distortion profile (Figure 7). A single

Algorithm 3 Reward Guided Refinement Loop

Require: Initial transcript T, baseline reward R_0, patient profile P, max iterations N

$best\_output \leftarrow T$
$best\_eval \leftarrow EvaluateSession(T)$
$best\_reward \leftarrow best\_eval.reward$
$target \leftarrow 1.3 \cdot R_0$
$i \leftarrow 1$
while $i \leq N$ **do**
  if $best\_reward \geq target$ **and** $i > 1$ **then**
    break ▷ Sufficient improvement achieved
  end if
  $feedback\_list \leftarrow [\,]$
  if $best\_eval.parts[distortion\_match] < 30.0$ **then**
    $feedback\_list \leftarrow$ *"Add more weighted examples."*
  end if
  if $feedback\_list$ = empty **then**
    break
  end if
  $feedback\_str \leftarrow$ Join $feedback\_list$
  $new\_eval \leftarrow EvaluateSession(new\_output)$
  $new\_output \leftarrow RefineSession(P, feedback\_str)$
  $new\_reward \leftarrow new\_eval.reward$
  if $new\_reward > best\_reward$ **then**
    $best\_output \leftarrow new\_output$
    $best\_eval \leftarrow new\_eval$
    $best\_reward \leftarrow new\_reward$
  else
    break ▷ No improvement, stop refinement
  end if
  $i \leftarrow i + 1$
end while
return $best\_output, best\_eval$

LLM (GPT-3.5-turbo[1]) then produces a 20–30 turn dialogue reflecting CBT techniques (e.g., Socratic questioning). We explicitly prompt it to embed distortions (catastrophizing, dichotomous thinking) according to the patient's weights.

Each transcript is evaluated via a custom scoring function (Algorithm 1), yielding a reward based on distortion fidelity and therapeutic quality. We use a 30% improvement threshold, feedback-based editing, and early stopping (Algorithm 3). A refinement-specific system role and prompt (Appendix D, Figures 9–11) are repeated for all patients. Appendix G shows a simulated session transcript with reward scores for each distortion.

4.3 Measuring Naturalism and Coherence in Patient Transcripts

We evaluate the quality of generated patient responses using a specialized LLM prompt that measures naturalism and coherence in a controlled, consistent format. This prompt (Appendix D, Figure 10) instructs the LLM to act as an expert evaluator trained in therapeutic communication and to assess responses on two dimensions: (1) *Naturalism*, which captures the realism, fluency, and emotional plausibility of the patient's speech; and (2) *Coherence*, which evaluates how well the response aligns with the preceding therapist statement in terms of contextual relevance and logical progression.

4.4 Evaluating Sessions using CoD Predispositions

As noted earlier, each patient is assigned a dictionary of distortion predispositions, representing the likelihood or intensity with which they exhibit specific CoDs (e.g., catastrophizing, emotional reasoning). Evaluation begins by iterating through a predefined set of regular expression patterns associated with each distortion type. These patterns are applied to the session transcript to identify and count linguistic indicators of each distortion. To quantify how well a patient's session aligns with their CoD profile, we define a normalized distortion match score. This score weights the frequency of distortion patterns in the session text by the patient's predisposition and normalizes it by text length and total distortion weights (Algorithm 4):

$$\mathrm{S} = \frac{1}{\left(\sum_{d \in D} w_d \cdot L\right)} \sum_{d \in D} \left(c_d \cdot (w_d + \varepsilon)\right) \quad (3)$$

where D is the set of cognitive distortions, c_d is the total raw match count for distortion d, w_d is the patient-specific predisposition weight for distortion d, ε is a small smoothing constant (e.g., 10^{-6}) to prevent zero-multiplication, and L is the number of words in the session text T.

To evaluate sessions and ensure alignment with each patient's profile, we define a distortion scoring function (Algorithm 4) that multiplies each distortion's raw frequency by its predisposition weight (adding a small constant for zero-weight). This yields a weighted distortion count showing how often and strongly that distortion appears. We also track the number of sentences containing distortions, then normalize by the total weighted sum and session word count for fair comparisons. The output includes raw/weighted counts, sentence-level spread, and the final normalized score—critical for session evaluation and refinement.

4.5 Refining Sessions with Few-Shot Learning

A key element of our architecture is a refinement function that iteratively improves therapy sessions

[1]Chosen for balanced quality, speed, and cost, although in general the proposed method is model-agnostic.

Algorithm 4 DistortionMatch

Require: Session text T, patient predisposition map P
Ensure: Normalized distortion score

Initialize $D_{counts}, D_{details} \leftarrow \{\}$
$total\_weighted \leftarrow 0$
$sum\_weights \leftarrow \sum P[d]$ or 1.0 if P is empty
while distortion d in $DISTORTION\_KEYWORDS$ **do**
  $weight \leftarrow P[d]$ or 1.0
  $raw\_count \leftarrow 0$
  $matches \leftarrow \emptyset$
  Split T into sentences S
  while $pattern \in$ `DISTORTION_KEYWORDS[d]` **do**
    for all $s_i \in S$ **do**
      if $pattern$ matches s_i **then**
        Add i to $matches$
      end if
    end for
    $raw\_count \leftarrow raw\_count + len(matches)$
  end while
  $weighted \leftarrow raw\_count \cdot (weight + \varepsilon)$
  $D_{counts}[d] \leftarrow weighted$
  $total\_weighted \leftarrow total\_weighted + weighted$
  $D_{details}[d] \leftarrow \{raw\_count, matches, weighted\}$
end while
$text\_length \leftarrow$ word count of T
$denominator \leftarrow (sum\_weights \cdot text\_length)$
$score \leftarrow total\_weighted / denominator$
return $\{D_{details}, D_{counts}, total\_weighted, score\}$

(Algorithm 3). It takes the current transcript, structured feedback, patient distortion weights, background, and notes from previous sessions, then revises the session. CoDs are embedded according to distortion weights, and the refinement prompts also include background/progress details (Figures 9 and 11). This feedback loop parallels RL policy refinement: rather than numeric gradients, qualitative metrics reshape prompts to steer improvement. Each revised session then returns a detailed transcript, distortion breakdown, reward metrics, and synthesized progress notes, laying groundwork for adaptive, interpretable, CBT-aligned AI systems.

4.6 Inline Annotation of Cognitive Distortions

To enhance model interpretability and learning, we synthetically generated inline annotations and instructed the LLM to incorporate the annotations within patient responses to explicitly mark cognitive distortions (e.g., "I always fail at everything [overgeneralization]"). Examples are presented in Figures 3 and 6. This approach serves as a form of weak supervision, providing the model with explicit semantic cues to associate linguistic patterns with their corresponding distortion categories. Our hypothesis was that this tagging strategy would improve classification performance by guiding the model's attention to distortion-relevant language.

Therapist	Hello, [Patient Name]. It's good to see you again. How have you been since our last session?
Patient	Hi, Doctor. I've been feeling a bit overwhelmed with [overgeneralization] all the responsibilities at home. It's like everything is always on my shoulders.
Therapist	It sounds like you've been experiencing a lot of pressure. Let's explore this feeling of everything being on your shoulders. Are there times when you do feel support from others?
Patient	Well, sometimes my family helps out, but I always feel like [should statements] I should be doing more. It's hard to let others take over.

Figure 3: Example of Inline Tagging

4.7 Generalization to Real-World Dialogues

We wanted to evaluate how well did our synthetic LLM-generated data generalized to real therapeutic interactions. We compared our generated sessions with anonymized examples from real therapist-patient datasets, including the Therapist QA corpus.[2] The LLM-generated dialogues—especially those refined through our reward-guided loop—closely matched real-world tone, structure, and cognitive patterns.

We also evaluated model performance on held-out real data. As expected, models trained solely on synthetic data underperformed, but hybrid models combining LLM and public data generalized well. This confirms that synthetic content can enhance generalization when paired with annotated examples. Tables 5 and 6 provide side-by-side comparison and similarity metrics, illustrating the linguistic and therapeutic realism of generated sessions.

4.8 Storing Optimized Session Data

As noted earlier, each simulated therapy session was iteratively refined and evaluated using a custom reward function, and only the highest-quality output, based on reward maximization, was retained for downstream use. At the end of the refinement loop, we stored a structured record for each session. The structured record includes a patient identifier, session number, progress notes, final reward score, and the full session transcript.

4.9 Dataset Variants and Classifiers

We use the public and anonymized dataset created by Shreevastava and Foltz (2021) composed of speeches that correspond to 10 types of "cog-

[2] https://www.kaggle.com/arnmaud/therapist-qa

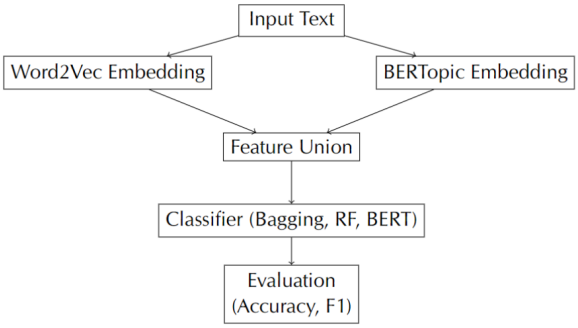

Figure 4: Pipeline for CoD Detection

nitive distortions" and neutral speeches categorized as "no distortion" type. This dataset contains 2530 annotated examples by experts; a summary is provided in Table 7. Our study evaluated models trained on the following: data without tags (labeled-no_inline_tags), data with inline tags (labeled-inline_tags), the public CoD dataset (labeled-public_cod), and a hybrid of LLM and public CoD data.

Although we initially experimented with Multilayer Perceptron, Decision Tree, Support Vector Machine, K-Nearest Neighbor, Gradient Boosting, Bagging, and Random Forest in our preliminary experiments due to their popularity in prior work with CBT techniques (Madububambachu et al., 2024; Lorenzoni et al., 2024), we ultimately included the three top-performing models (Gradient Boosting, Bagging, and Random Forest) in our full evaluation.

We built a pipeline using Word2Vec (Mikolov et al., 2013) and BERTopic (Grootendorst, 2022) for richer CoD representations, combining lexical and topical clues (Figure 4). Similar approaches appear in Alhaj et al. (2022) for Arabic CoD detection, Sharma and Sirts (2024) for depression markers, and Kellert and Mahmud Uz Zaman (2022) for semantic insight beyond literal meanings. Akash and Chang (2024) use BERTopic on expanded short texts, and Vanin et al. (2024) analyze therapist remarks for deeper discourse patterns.

We retrain Word2Vec on the generated CoD corpus for sentence-level embeddings and employ BERTopic to yield topic-based labels for each sample. By merging them via a `FeatureUnion`, Word2Vec captures fine-grained semantics while BERTopic encodes higher-level thematic structure. This hybrid embedding strategy enhances classification robustness on CoD tasks.

Training Dataset	Avg. Acc.	Best Acc.	Best Algorithm
`LLM w/no tags`	0.45	0.62	Gradient Boosting
`LLM w/tags`	0.51	0.65	Bagging
`Annotated public dataset`	0.68	0.73	Bagging
`LLM w/no tags, 30% public`	0.70	0.75	Random Forest
`LLM w/no tags, 70% public`	0.83	0.85	Bagging
`LLM w/tags, 30% public`	0.71	0.76	Bagging
`LLM w/tags, 70% public`	0.84	0.86	Random Forest

Table 1: Accuracies for different training approaches.

5 Results

The results presented in Table 1 reflect a comprehensive evaluation of models trained for CoD classification using varied dataset configurations and algorithms, including models trained solely on LLM-generated data, with and without inline CoD tags and/or a public dataset, and hybrid combinations. The trends provide valuable insights into how annotation strategies and data provenance affect classification performance. Performance improvements were observed across all models, particularly when inline tagging was included.

5.1 Baseline Observations

Classification performance was evaluated using accuracy and weighted F1. Training classifiers on only the LLM generated datasets, and testing them on the annotated public dataset, we achieved a best weighted F1=0.63 (Table 2). When we trained classifiers only on the annotated dataset, we achieved an average weighted F1=0.70. This relatively modest performance suggests that while LLMs are capable of generating realistic therapy data, the lack of explicit signals confirmed in ground truth makes it harder for conventional classifiers to learn effective decision boundaries. When we combined the generated dataset with ground truth data, there was marked improvement across the board.

Training on LLM-generated data with no inline tagging and testing on the annotated public dataset yielded an average F1=0.58. To test our hypothesis that inline tagging should improve precision, we found that introducing inline CoD tags (e.g., `[catastrophizing]` yielded a measurable improvement, with average F1 rising to 0.60 with

Training Dataset	Avg F1	Best F1	Best Algorithm
LLM w/no tags	0.52	0.58	Gradient Boosting
LLM w/tags	0.60	0.63	MLP
Annotated public dataset	0.67	0.70	Bagging
LLM w/no tags, 30% public	0.70	0.73	Bagging
LLM w/no tags, 70% public	0.84	0.85	Bagging
LLM w/tags, 30% public	0.73	0.74	Bagging
LLM w/tags, 70% public	0.84	0.86	Random Forest

Table 2: Weighted F1-scores for Binary CoD detection.

Data	Bagging	P	R	F1
Public	Class 0	0.79	0.37	0.50
	Class 1	0.72	0.94	0.82
	Weighted	0.74	0.73	0.70
LLM w/tags, 30% public	Class 0	0.82	0.44	0.58
	Class 1	0.74	0.94	0.83
	Weighted	0.77	0.76	0.74
LLM w/tags, 70% public	Class 0	0.83	0.77	0.80
	Class 1	0.87	0.91	0.89
	Weighted	0.86	0.86	0.86

Table 3: Results (Bagging, varying data conditions).

a best-case performance of 0.63. Average and best accuracy also rose, confirming our hypothesis that weak supervision via inline annotation provides useful inductive bias, helping models learn class boundaries more effectively by tying linguistic patterns directly to distortion labels.

5.2 Evaluating Performance with Inline Tagging

We also studied the combination of inline tagged LLM data with scaled proportions of the annotated public dataset, and found marked improvements. We found reasonably strong performance when we combined the LLM data with 30% of the ground truth, achieving a best accuracy=0.76 and best weighted F1=0.74. As expected when we scaled the proportion of the ground truth in the training dataset to 70%, the performance increased.

Notably, the human-annotated public dataset, when combined with the tagged LLM-generated data, outperformed every other model. This underscores the importance of human-annotated and field-tested corpora, likely due to their greater consistency and alignment with diagnostic standards. Our results also show that inline CoD tagging both benefits models when used alone and also amplifies learning when mixed with human-validated data. The result reflects an optimal balance between synthetic diversity and supervised precision.

5.3 Ablation Studies

To validate our approach, we performed several ablation studies. First, removing inline tags from the hybrid model led to a notable performance drop. Training on only synthetic data (with and without tags) confirmed a dependency on human-labeled data (Tables 1 and 2). Models trained without public data consistently underperformed.

We also varied the proportion of public data in hybrid setups (30% vs. 70%) to study scaling effects. With 30% public data and untagged LLM examples, Random Forest reached average F1=0.70 (max 0.73). Adding inline tags improved average F1 to 0.73 (max 0.74). Increasing public data to 70% further improved all metrics, highlighting the value of larger annotated corpora.

These results show that LLM-generated content enhances generalization when paired with ground truth data. The best performance (average F1=0.73, max F1=0.74) was achieved on tagged LLM data combined with public data, confirming the benefits of hybrid training and inline annotation.

5.4 Evaluating Bagging

Bagging shows the best overall performance on the public dataset, revealing key insights into model performance on imbalanced CoD data (see Table 3). While the overall accuracy is 73%, accuracy masks uneven class performance. Class 1 (distortion-present cases) achieves strong results with an F1=0.82 and recall=0.94, indicating the model correctly identifies most distortion cases. However, performance on Class 0 (distortion-absent) is weaker, with an F1=0.50 and recall=0.37. The weighted F1=0.70 offers a more balanced view, accounting for class proportions and performance. These findings highlight that although the model appears accurate overall, it struggles with the minority class. Future improvements could include rebalancing strategies or class-specific tuning to reduce this performance gap.

Combining tagged LLM generated data with 30% of the public dataset, we find that Bagging achieved an overall accuracy of 0.76. A closer

Data	BERT	P	R	F1
Public	Class 0	1.00	0.00	0.01
	Class 1	0.63	1.00	0.77
	Weighted	0.77	0.63	0.49
LLM w/tags, 30% public	Class 0	0.76	0.18	0.29
	Class 1	0.67	0.97	0.79
	Weighted	0.70	0.68	0.61
LLM w/tags, 70% public	Class 0	0.68	0.65	0.66
	Class 1	0.80	0.82	0.81
	Weighted	0.75	0.76	0.76

Table 4: Results (BERT, varying data conditions).

look at class-specific metrics again reveals a performance imbalance. Class 1 (distortion-present) was well-detected, with a recall=0.94 and F1=0.83, showing strong sensitivity to identifying cognitive distortions. However, the model struggled with Class 0 (distortion-absent), achieving recall=0.44 and an F1=0.58, indicating a tendency to misclassify non-distorted responses. The weighted F1=0.74 offers a more reliable summary. While Bagging performs well on the dominant class, improvements are needed to boost recall and precision on the minority class to reduce false positives and ensure balanced detection across categories.

5.5 BERT Comparison

Although widely used in NLP, our experiments show BERT underperforms in low-data settings compared to classical models (Table 4). On the public dataset, BERT achieves 63% accuracy, but this hides extreme class imbalance: Class 1 (distortion-present) achieves perfect recall (1.00) and F1=0.77, while Class 0 (distortion-absent) is completely misclassified (F1=0.01, recall=0.00). The model defaults to predicting distortions, inflating macro precision (0.82), but macro and weighted F1 (0.39, 0.49) reflect poor generalization. This highlights the need for better-balanced data or domain-specific fine-tuning.

Using a hybrid dataset (tagged LLM + 30% public), accuracy improves to 68%, and Class 1 maintains high recall (0.97) and F1=0.79. However, Class 0 remains weak (F1=0.29, recall=0.18), indicating persistent false positives. The weighted F1-score of 0.61 shows improved generalization, but class bias remains. While LLM data helps, it alone doesn't fully resolve BERT's skew.

With 70% public data and tagged LLM samples, BERT achieves its best balance: 76% accuracy and improved performance on both classes. This suggests that incorporating more high-quality labeled data allows BERT to generalize better and approach Bagging performance, reducing prior bias.

5.6 Class Imbalance

Given the imbalance between classes, we use PR curves as a more informative diagnostic tool than ROC curves. These curves visualize the trade-off between precision and recall across different thresholds and help identify models that retain high precision at varying levels of recall. To visualize model strengths and weaknesses, we include class-specific PR curves and confusion matrices in the Appendix (Figure 12). These extended diagnostics reveal, for instance, that while BERT achieves high recall on distortion-present cases, it underperforms in detecting distortion-absent responses, validating a pattern also seen in Bagging and Random Forest classifiers. This supports our argument that distortion-absent detection remains a challenging and underrepresented area for improvement.

6 Conclusion

This study introduces a comprehensive framework for generating, annotating, and classifying CoDs using LLMs, simulated patient profiles, and targeted reinforcement. By leveraging LLM-generated therapy dialogues guided by structured CoD predispositions, we created a synthetic dataset with realistic clinical dynamics. A reinforcement learning-inspired refinement loop, driven by a reward function balancing naturalism, coherence, and distortion alignment, was used to iteratively improve each session's quality and distortion fidelity.

We integrated hybrid feature engineering, combining Word2Vec embeddings for lexical-semantic details with BERTopic features for high-level thematic structure. This dual approach helps classifiers learn from micro-level distortions and macro-level discourse cues. Empirical results show that models trained on hybrid data (inline tagging plus public datasets) outperform baselines, particularly in precision and F1 for distortion-present classes.

Despite improvements, the pipeline struggles with distortion-absent samples, underscoring the need for class rebalancing and richer supervision. Overall, it provides a scalable, clinically informed approach to modeling cognitive distortions. The pipeline and generated data will be released publicly to encourage replication and broader adoption.

7 Ethical Considerations

This work involves the generation of synthetic therapist–patient dialogues using LLMs, which raises important ethical considerations. First, although the patient profiles are entirely fictional and devoid of real personal data, care was taken to avoid reinforcing stereotypes related to mental health conditions, age, or identity. Each virtual patient was designed to simulate cognitive distortion patterns realistically but respectfully.

Second, the use of LLMs to simulate mental health interactions should not be interpreted as a replacement for licensed clinical professionals. While this research aims to support therapeutic NLP applications and cognitive distortion detection, its outputs are not intended for diagnostic or treatment use. The models presented here are strictly research tools for studying patterns in text, not decision-making systems for clinical practice. Lastly, transparency and reproducibility were prioritized by releasing both the code and datasets. However, researchers using this data must ensure that downstream applications maintain ethical boundaries, especially in high-stakes domains such as mental health and education.

8 Limitations & Future Work

While the proposed approach is promising, several limitations remain. First, LLM-generated patient responses, though diverse in cognitive distortions, may lack the nuance of real-world therapy, limiting generalizability. Future work could incorporate Mixture-of-Personas (MoP) prompting to better simulate patient variation, as demonstrated by Harel-Canada et al. (2024), and integrate moral-cultural context to enhance cross-cultural sensitivity and distortion modeling, following the framework of Ramezani and Xu (2023). These strategies may also improve semantic alignment with real dialogues.

Second, our reward-guided reinforcement loop relies on heuristic-based scoring (e.g., distortion and naturalism), which may bias optimization. Third, BERT's underperformance on distortion-absent cases suggests that even large models need domain-specific tuning and balanced training. Lastly, the inline tagging strategy improves model learning but may introduce artifacts not present in natural conversation. Future work should explore more subtle, latent methods of guiding model attention without altering patient speech explicitly.

References

Pritom Saha Akash and Kevin Chen-Chuan Chang. 2024. Enhancing short-text topic modeling with llm-driven context expansion and prefix-tuned vaes. *Preprint*, arXiv:2410.03071.

Fatima Alhaj, Ali Al-Haj, Ahmad Sharieh, and Riad Jabri. 2022. Improving arabic cognitive distortion classification in twitter using bertopic. *International Journal of Advanced Computer Science and Applications*, 13(1).

Janire Arana, Mikel Idoyaga, Maitane Urruela, Elisa Espina, Aitziber Atutxa Salazar, and Koldo Gojenola. 2024. A virtual patient dialogue system based on question-answering on clinical records. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 2017–2027, Torino, Italia. ELRA and ICCL.

Santiago Berrezueta-Guzman, Mohanad Kandil, María-Luisa Martín-Ruiz, Iván Pau de la Cruz, and Stephan Krusche. 2024. Future of adhd care: Evaluating the efficacy of chatgpt in therapy enhancement. *Healthcare*, 12(6).

Zhiyu Chen, Yujie Lu, and William Yang Wang. 2023. Empowering psychotherapy with large language models: Cognitive distortion detection through diagnosis of thought prompting. *Preprint*, arXiv:2310.07146.

Ignacio de Toledo Rodriguez, Giancarlo Salton, and Robert Ross. 2021. Formulating automated responses to cognitive distortions for CBT interactions. In *Proceedings of the 4th International Conference on Natural Language and Speech Processing (ICNLSP 2021)*, pages 108–116, Trento, Italy. Association for Computational Linguistics.

Maarten Grootendorst. 2022. Bertopic: Neural topic modeling with a class-based tf-idf procedure. *Preprint*, arXiv:2203.05794.

Fabrice Harel-Canada, Hanyu Zhou, Sreya Muppalla, Zeynep Yildiz, Miryung Kim, Amit Sahai, and Nanyun Peng. 2024. Measuring psychological depth in language models. *Preprint*, arXiv:2406.12680.

Olga Kellert and Md Mahmud Uz Zaman. 2022. Using neural topic models to track context shifts of words: a case study of COVID-related terms before and after the lockdown in April 2020. In *Proceedings of the 3rd Workshop on Computational Approaches to Historical Language Change*, pages 131–139, Dublin, Ireland. Association for Computational Linguistics.

Sehee Lim, Yejin Kim, Chi-Hyun Choi, Jy-yong Sohn, and Byung-Hoon Kim. 2024. ERD: A framework for improving LLM reasoning for cognitive distortion classification. In *Proceedings of the 6th Clinical Natural Language Processing Workshop*, pages 292–300, Mexico City, Mexico. Association for Computational Linguistics.

Zhengyuan Liu, Siti Salleh, Pavitra Krishnaswamy, and Nancy Chen. 2024. Context aggregation with topic-focused summarization for personalized medical dialogue generation. In *Proceedings of the 6th Clinical Natural Language Processing Workshop*, pages 310–321, Mexico City, Mexico. Association for Computational Linguistics.

Giuliano Lorenzoni, Cristina Tavares, Nathalia Nascimento, Paulo Alencar, and Donald Cowan. 2024. Assessing ml classification algorithms and nlp techniques for depression detection: An experimental case study. *Preprint*, arXiv:2404.04284.

Kevin Lybarger, Justin Tauscher, Xiruo Ding, Dror Benzeev, and Trevor Cohen. 2022. Identifying distorted thinking in patient-therapist text message exchanges by leveraging dynamic multi-turn context. In *Proceedings of the Eighth Workshop on Computational Linguistics and Clinical Psychology*, pages 126–136, Seattle, USA. Association for Computational Linguistics.

Mounica Maddela, Megan Ung, Jing Xu, Andrea Madotto, Heather Foran, and Y-Lan Boureau. 2023. Training models to generate, recognize, and reframe unhelpful thoughts. *Preprint*, arXiv:2307.02768.

U. Madububambachu, A. Ukpebor, and U. Ihezue. 2024. Machine learning techniques to predict mental health diagnoses: A systematic literature review. *Clinical Practice & Epidemiology in Mental Health*, 20:e17450179315688.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *Preprint*, arXiv:1301.3781.

Hongbin Na. 2024. CBT-LLM: A Chinese large language model for cognitive behavioral therapy-based mental health question answering. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 2930–2940, Torino, Italia. ELRA and ICCL.

Emma Pirnay. 2023-04-27. We spoke to people who started using chatgpt as their therapist. *Vice*.

Paolo Raile. 2024. The usefulness of chatgpt for psychotherapists and patients. *Humanities and Social Sciences Communications*, 11:1–8.

Aida Ramezani and Yang Xu. 2023. Knowledge of cultural moral norms in large language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 428–446, Toronto, Canada. Association for Computational Linguistics.

Neha Sharma and Kairit Sirts. 2024. Context is important in depressive language: A study of the interaction between the sentiments and linguistic markers in Reddit discussions. In *Proceedings of the 14th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, pages 344–361, Bangkok, Thailand. Association for Computational Linguistics.

Sagarika Shreevastava and Peter Foltz. 2021. Detecting cognitive distortions from patient-therapist interactions. In *Proceedings of the Seventh Workshop on Computational Linguistics and Clinical Psychology: Improving Access*, pages 151–158, Online. Association for Computational Linguistics.

Gopendra Vikram Singh, Sai Vardhan Vemulapalli, Mauajama Firdaus, and Asif Ekbal. 2024. Deciphering cognitive distortions in patient-doctor mental health conversations: A multimodal LLM-based detection and reasoning framework. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 22546–22570, Miami, Florida, USA. Association for Computational Linguistics.

Alexander Vanin, Vadim Bolshev, and Anastasia Panfilova. 2024. Applying llm and topic modelling in psychotherapeutic contexts. *Preprint*, arXiv:2412.17449.

Bichen Wang, Pengfei Deng, Yanyan Zhao, and Bing Qin. 2023. C2D2 dataset: A resource for the cognitive distortion analysis and its impact on mental health. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10149–10160, Singapore. Association for Computational Linguistics.

Ruiyi Wang, Stephanie Milani, Jamie C. Chiu, Jiayin Zhi, Shaun M. Eack, Travis Labrum, Samuel M Murphy, Nev Jones, Kate V Hardy, Hong Shen, Fei Fang, and Zhiyu Chen. 2024. PATIENT-ψ: Using large language models to simulate patients for training mental health professionals. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 12772–12797, Miami, Florida, USA. Association for Computational Linguistics.

A Cognitive Distortions

Overgeneralization: Drawing broad, sweeping conclusions based on one event or limited evidence. `I failed once, so I'll always fail, always, never, everyone, nobody, everything`

All-or-Nothing Thinking: Viewing situations in absolute, black-or-white terms without acknowledging nuance or middle ground. `If it's not perfect, it's a total failure, completely, totally, entirely, ruined, worthless`

Catastrophizing: Expecting the worst possible outcome or exaggerating the consequences

of a situation. `If I make one mistake, I'll lose everything, disaster, catastrophe, horrible, the worst possible outcome`

Mind Reading: Assuming the thoughts, motives, or judgments of others without evidence. `they think I'm incompetent, they must hate me, everyone can see I am nervous`

Should Statements: Using rigid rules or expectations to guide behavior or evaluate others, often resulting in guilt, frustration, or resentment `I should always be in control, I should, I must, I have to, I ought to`

Disqualifying the Positive: Dismissing or minimizing positive experiences or feedback by attributing them to luck or external factors. `They praised me, but they didn't mean it, It doesn't count, it was just luck, anyone could have done that`

Personalization: Blaming oneself for events outside of one's control, or interpreting external events as being directly related to oneself. `It's my fault the meeting went badly, It's all my fault, I caused this, I am to blame`

Mental Filter: Focusing exclusively on the negative details of a situation while ignoring any positives. `I got one critical comment, so the entire presentation was a disaster, Only focus on the negative, nothing good ever happens`

Labeling: Assigning a fixed, global label to oneself or others based on specific behaviors. `I'm a failure, He's useless, I'm a failure, I'm worthless, they are nothing but a loser`

B Full Sample Therapy Session

Therapist: Good morning. I'm Dr. Morgan, and I'm here to help you today. How have you been feeling recently?
Patient: Well, to be honest, I always feel overwhelmed. I never [overgeneralization] have time for myself because everyone around me always needs something.
Therapist: It sounds like you're feeling like you're constantly stretched thin, like you have to do everything for everyone. Do you think there might be room for some balance between taking care of others and taking care of yourself?
Patient: I'm not sure. I feel like if I'm not there for everyone, then everything [should statements] will fall apart. It's like if I'm not perfect at caregiving, then I'm completely useless.
Therapist: It's understandable to feel that pressure, but is it possible that expecting perfection in caregiving might be setting an unrealistic standard for yourself? Could there be some middle ground where you're doing your best without having to be perfect?
Patient: I guess so, but I can't help catastrophizing. I always think that if [personalization] I make a mistake, it will lead to a disaster and everything will crumble.
Therapist: Catastrophizing can make situations seem much worse than they are. What evidence can we find that supports the idea that a mistake you make will truly lead to a catastrophe?
Patient: I just feel like everything is on the line all the time. It's like if I don't do everything right, then nothing will work out for me or for those I care for.
Therapist: That sounds very challenging to carry that weight. Let's explore the possibility that maybe not everything is as catastrophic as it seems. Could there be some room for things to not be perfect and still turn out okay?
Patient: I should be able to handle everything perfectly though. I must make sure everyone is taken care of, or else I'll let them down.
Therapist: It seems like there's a lot of pressure on you to meet these "should" statements. What would happen if you were to let go of some of these unrealistic expectations and allow for imperfection?
Patient: But if I don't do everything right, then I feel like it doesn't count. Any small successes I have feel insignificant compared to what I think I should achieve.
Therapist: It sounds like you're focusing on the negatives and discounting the positives. Is there a way to acknowledge your achievements,

no matter how small, without disqualifying them?
Patient: It's just hard not to take everything personally. I feel like if something goes wrong, it's all my fault. I caused it, and everyone suffers because of me.
Therapist: Personalizing situations can lead to an overwhelming sense of responsibility. Is it possible that other factors beyond your control could contribute to challenges, and that not everything is solely your fault?
Patient: I know what people think of me. I can tell when they're disappointed or judging me. It's like they all see me as a failure.
Therapist: It seems like you're engaging in mind reading, assuming you know what others are thinking. It can feel isolating when we assume others are thinking negatively of us. What if we paused and asked ourselves, "Do I have solid evidence that someone sees me as a failure, or am I interpreting it that way?" Sometimes, we don't truly know what another person is thinking.
Patient: I guess I don't really know for sure. But it's so easy to interpret a look or a tone of voice as disappointment.
Therapist: Absolutely—it's natural to worry about others' judgments. But you might ask, "Could there be another explanation?" Maybe that look is just stress they feel about their own problems. It doesn't always have to mean you did something wrong.
Patient: I see your point. I can't really prove they're judging me.
Therapist: Exactly. And perhaps challenging that assumption frees up emotional space. Rather than mind reading, can we focus on what's actually happening in a situation? That can be more grounding and less stressful. Let's think about how you could practice noticing those automatic thoughts—like assuming someone's disappointed—and then testing them.
Patient: Maybe next time, I could check in with the person or at least tell myself, "I don't have evidence for this; they could just be busy."
Therapist: That's a great step. Over time, challenging mind reading can help you see that most situations are more neutral than they initially feel. Are you willing to try that this week—keeping track of moments you feel judged, then asking if there's real proof?
Patient: Yes, I can try. I guess it can't hurt to be more fair to myself.
Therapist: Being fair to yourself is important. Let's recap: we've identified several unhelpful thinking patterns today. The plan is to notice and gently question these thoughts. We'll see how that changes how you feel. Does that sound good?
Patient: It does. I'll try to be more mindful and maybe write my thoughts down.
Therapist: Excellent. That's a great strategy. We can review next time how this approach affects your stress levels and whether it eases the pressure you've been feeling.

C Virtual Patient Profile Attributes

Simulated patient profiles varied in terms of age distribution, background, and CoD weighting. We elaborate on each of these factors below:

- **Age Distribution:** Patients were assigned ages ranging from 12 to 80 years, reflecting a wide demographic relevant to both adolescent and adult therapy contexts. This range ensures the inclusion of diverse emotional development stages and therapeutic needs.
- **Background Descriptions:** Each patient has a one-sentence background tailored to evoke plausible therapeutic dialogue. These were manually crafted using a set of archetypes drawn from real-world therapy research (e.g., trauma survivors, high-performing professionals under stress, caregivers, retirees, veterans). The goal was to mirror varied emotional contexts and lived experiences without introducing clinical diagnoses.
- **Distortion Weighting:** Each profile contains a dictionary of cognitive distortions with numerical weights (0–10) representing the likelihood or intensity of that distortion manifesting in their dialogue. These values were assigned using stratified random sampling. As alluded to earlier, only 33% of the virtual patients were assigned weighted distortions using a uniform sampling strategy bounded by thematic coherence in their background.

Balancing manual curation and randomized variation allows for enhanced realism and diversity,

prompt	Generate a diverse set of 100 fictional patient profiles for use in simulating cognitive behavioral therapy (CBT) sessions. Each profile should include a unique patient ID, name (e.g., "Patient17"), age (ranging from 18 to 80), and a one-paragraph background describing the patient's life context in a respectful and realistic manner. Avoid any references to race, gender, religion, or specific cultural identities. Ensure that backgrounds reflect a wide variety of life experiences (e.g., students, retirees, caregivers, professionals, etc.) without reinforcing stereotypes about mental health or age. Each profile should also contain a '$primary_cod$' dictionary that lists weights (0–10) for the following cognitive distortions overgeneralization, mind reading, personalization, catastrophizing, all-or-nothing thinking, mental filter, disqualifying the positive, emotional reasoning and should statements. Exactly 33% of the patients should exhibit cognitive distortions (i.e., non-zero weights), and the remaining 67% should have zero across all distortion types. Assign distortion weights realistically and sparingly, with no more than 3–4 elevated distortions per distorted profile. Ensure the profiles sound plausible and do not exaggerate or caricature emotional states. These profiles are for use in training AI models to detect patterns of thought—not diagnoses—and should model everyday human challenges with empathy and care.

Figure 5: LLM Prompt for Virtual Patient Profiles

with manual curation enabling human judgment in narrative plausibility and randomized distortion increasing generalizability and avoiding overfitting to a narrow emotional schema. A sample session output is shown in Appendix G.

D Extended Prompts

The full prompt used to generate each virtual patient profile is provided in Figure 5, incorporating the profile attributes described in Appendix C. This prompt establishes the foundational characteristics of each simulated patient, including age, background, and a weighted cognitive distortion profile. These components ensure diversity and realism while preserving ethical boundaries and avoiding stereotypical representations.

To simulate therapeutic interactions, we use a dedicated system prompt that assigns the LLM the role of a CBT-trained senior therapist conducting a session with the virtual patient (Figure 7). This

Therapist	Good afternoon, how have you been since our last session?
Patient	Oh, Doctor, everything is just falling apart. [overgeneralization] I always mess things up, and now I'm ruined.
Therapist	It sounds like you're feeling overwhelmed with everything going on. Let's break it down. Can you tell me more about what specifically has been challenging for you?
Patient	Well, I lost another big contract last week. It's like nothing [mental filter] ever works out for me, and I'm starting to believe I'm just worthless.

Figure 6: Another Example of Inline Tagging

role	You are simulating a therapy session between a CBT-trained senior therapist and $patient\_name$. This is Session $session\_num$ out of $NUM\_SESSIONS$ total sessions. Use text from $DISTORTION\_KEYWORDS$ to generate patient responses.

Figure 7: CBT Prompt

prompt emphasizes therapeutic tone and conversational structure grounded in established CBT communication patterns.

Since our approach uses few-shot prompting to generate multi-turn therapy sessions, each session prompt is dynamically constructed to include the patient's background, their individualized cognitive distortion profile, and progress notes from the previous session. This structure promotes continuity and simulates the feel of an ongoing therapeutic relationship. The template used to generate these session prompts is illustrated in Figure 8. Following the initial session generation, the system further constructs a refinement-specific system role and a contextualized refinement prompt that incorporates evaluator feedback to guide improvement. These components are shown in Figures 9 and 11, respectively.

To evaluate the quality of generated patient responses, we use a specialized evaluation prompt (Figure 10). This prompt instructs the LLM to act as an expert evaluator trained in therapeutic communication and to assess each session response along two dimensions: *naturalism*, measuring how human-like and contextually fluent the response is, and *coherence*, evaluating its logical alignment with the ongoing session narrative. These evaluation scores feed directly into the reward function guiding the reinforcement loop.

base prompt	Patient Distortions MUST be highlighted in the generated response, based on predispositions: *predisposition_dict*.
Background	*patient_background*
Progress	*previous_notes*
Instruction 1	Write 20-30 turns (Therapist: / Patient:) realistically.
Instruction 2	Show the patient using the above distortions as per the predisposition weights, but reflect some improvement if it's beyond session 1.
Instruction 3	Keep the conversation reflective of a CBT approach.
Instruction 4	The therapist is very experienced and uses CBT techniques (e.g., Socratic questioning, challenging negative thoughts).

Figure 8: Prompt design for API-access LLMs

role	Below is the original therapy session draft, followed by feedback that you must address. Revise the text SUBSTANTIALLY to incorporate the feedback using words from $DISTORTION\_KEYWORDS$.

Figure 9: Refinement System Role

Transcript	llm_output.
Prompt	Given a turn-by-turn therapist-patient exchange, your task is to assign two numeric scores to the patient's response using the criteria below. Output must strictly follow the format provided at the end.
Naturalism	Score based on how realistically and human-like the patient's response sounds. Consider fluency and grammaticality, appropriateness of tone in a therapeutic setting, whether it mimics how real patients might speak
Scoring Guide	1.0: Completely natural, indistinguishable from a real patient. Assign an appropriate number if somewhat stilted, but plausible. 0.0: Robotic, unnatural, or unrealistic in tone or phrasing.
Coherence	Score based on how well the patient's response connects to the therapist's previous statement. Consider relevance and logical progression, whether the patient is responding to the question or prompt, topic continuity and contextual alignment.
Scoring Guide	1.0: Fully coherent and contextually appropriate. Assign an appropriate number if partially relevant or mildly off-topic. 0.0: Disconnected, contradictory, or non-responsive.
Output Format	(strictly adhere to this): Naturalism: <float between 0.0 and 1.0> Coherence: <float between 0.0 and 1.0>

Figure 10: Naturalism and Coherence Scoring Prompt

Revised prompt	original session: llm_output.
Feedback	$feedback$ Patient Distortions to highlight (based on predispositions): $predisposition\_dict$.
Background	$background$
Progress	$previous\_notes$
Instruction	Please provide a revised session that addresses the above points.

Figure 11: Refinement System prompt

E PR Curves for CoD Detection

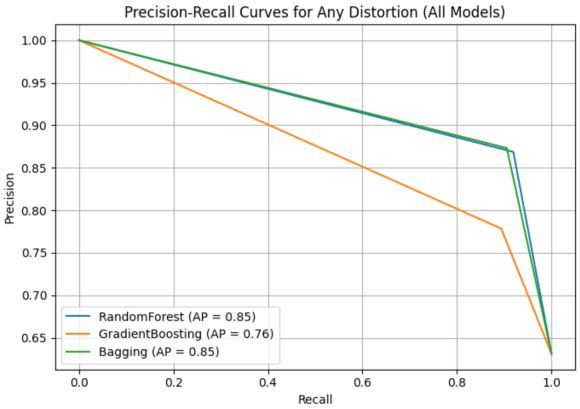

Figure 12: Precision-recall curves for models used in the study.

F Comparison with Real-World Dialogues

Real Dialogue (Therapist QA)	Generated Dialogue (LLM)
Patient: I ask her what was wrong and she replied: I hear voices in my ears but I dont see the people saying it. She says it happened during school doing a reading circle.	**Patient:** It's just hard not to take everything personally. I feel like if something goes wrong, it's all my fault. I caused it, and everyone suffers because of me.
Patient: She thought someone called her stupid and let the teacher know. The teacher said no one said anything.	**Patient:** I just feel like everything is on the line all the time. It's like if I don't do everything right, then nothing will work out for me or for those I care for.

Table 5: Comparison of real and LLM-generated therapist-patient dialogues. The generated session mirrors structure, tone, and cognitive distortion patterns observed in real-world interactions.

Metric	Score	Interpretation
BERTScore (F1)	0.70	Moderate to strong semantic similarity
BLEURT	0.16	Moderate semantic similarity
SBERT Cosine Similarity	0.51	Moderate semantic alignment in embedding space

Table 6: Similarity metrics comparing real and LLM-generated therapy dialogue. We achieve moderate alignment in the present study.

G Reward Weighted Session Outputs

Listing 2: Sample Session Output


```
{
    "patient_id": 2,
    "session_number": 1,
    "reward": 147.00001999999998,
    "progress_notes": "In this session,
        we achieved a final reward
        score of 147.00. The patient's
        dialogue contained signs of:
        overgeneralization,
        all-or-nothing thinking,
        catastrophizing, should
        statements, personalization,
        mental filter, labeling. Moving
        forward, we will continue
        addressing these distortions
        and track improvements.",
    "transcript": <see Appendix B>,
    "evaluation_scores": {
      "distortion_match": 147
          .00001999999998
    },
    "distortion_breakdown": {
      "overgeneralization": 108.000012,
      "all-or-nothing thinking": 4
          .000001,
      "catastrophizing": 1.000001,
      "mind reading": 0.0,
      "should statements": 15.000003,
      "disqualifying the positive": 0.0,
      "personalization": 10.000001,
      "mental filter": 8.000001,
      "labeling": 1.000001
    },
    "distortion_details": {
      "overgeneralization": {
        "raw_count": 12,
        "weight": 9.000001,
        "matched_sentence_count": 10,
        "weighted_count": 108.000012
      },
      "all-or-nothing thinking": {
        "raw_count": 1,
        "weight": 4.000001,
        "matched_sentence_count": 1,
        "weighted_count": 4.000001
      },
      "catastrophizing": {
        "raw_count": 1,
        "weight": 1.000001,
        "matched_sentence_count": 1,
        "weighted_count": 1.000001
      },
      "mind reading": {
        "raw_count": 0,
        "weight": 2.000001,
        "matched_sentence_count": 0,
        "weighted_count": 0.0
      },
      "should statements": {
        "raw_count": 3,
        "weight": 5.000001,
        "matched_sentence_count": 3,
        "weighted_count": 15.000003
      },
      "disqualifying the positive": {
        "raw_count": 0,
        "weight": 1.000001,
        "matched_sentence_count": 0,
        "weighted_count": 0.0
      },
      "personalization": {
        "raw_count": 1,
        "weight": 10.000001,
        "matched_sentence_count": 1,
        "weighted_count": 10.000001
      },
      "mental filter": {
        "raw_count": 1,
        "weight": 8.000001,
        "matched_sentence_count": 1,
        "weighted_count": 8.000001
      },
      "labeling": {
        "raw_count": 1,
        "weight": 1.000001,
        "matched_sentence_count": 1,
        "weighted_count": 1.000001
      }
    }
}
```


H Public Dataset

Distortion Type	Count
All-or-nothing thinking	100
Emotional Reasoning	134
Fortune-telling	143
Labeling	165
Magnification	195
Mental filter	122
Mind Reading	239
Overgeneralization	239
Personalization	153
Should statements	107
No Distortion	933
In Total	**2530**

Table 7: Details of the dataset used in this paper

I Computational Experiments

Model size, compute budget, and infrastructure: We used GPT-3.5-turbo via API for session generation and refinement. GPT-3.5 has approximately 6.7 billion parameters. While we did not train or fine-tune LLMs, generation involved approximately 6,000 prompt-response cycles over multiple sessions. All generation and scoring tasks were performed using OpenAI's hosted infrastructure. Classification experiments were conducted using `scikit-learn` and BERT-base (110M parameters) on a local workstation with an NVIDIA RTX 3090 GPU (24GB), 16GB RAM, and an Intel(R) Core(TM) Ultra 7 155H 1.40 GHz processor. Total compute time for classification training and evaluation was under 8 GPU hours.

Experimental setup and hyperparameters: We used a stratified train-test split (80/20) with 5-fold cross-validation where applicable. Models tested included Bagging, Random Forest, Gradient Boosting, and MLP. Hyperparameters were manually tuned based on validation performance. The best-found settings included:

> **Random Forest**: min_samples_split=10, max_depth=20.
>
> **Bagging**: n_estimators=50, base estimator = DecisionTreeClassifier().
>
> **Gradient Boosting**: learning_rate=0.1, n_estimators=50.

Descriptive statistics and result reporting: All reported results include both average and best-case performance across 3 random seeds. For classification metrics, we report accuracy and weighted F1-scores. Confusion matrices and class-specific metrics are provided for Bagging and BERT. PR curves for all major models are included in the appendix. Test set results were separated from validation data, and we clearly specify when results are derived from a single run or averaged.

External packages and implementations: We used the following libraries:

- `scikit-learn` (v1.3.2) for ML models and classification metrics.
- `BERTopic` (v0.15.0) for topic modeling.
- `nltk` (v3.8.1) for tokenization and lexical preprocessing.
- `gensim` (v4.3.1) for Word2Vec embeddings.
- OpenAI API for generation using GPT-3.5-turbo.

No modifications were made to external libraries. All preprocessing steps (tokenization, inline tagging, embedding generation) are described in the Methodology section.