Deciphering Cognitive Distortions in Patient-Doctor Mental Health Conversations: A Multimodal LLM-Based Detection and Reasoning Framework

Gopendra Vikram Singh[1], **Sai Vardhan Vemulapalli**[1], **Mauajama Firdaus**[2], **Asif Ekbal**[3],
[1]Department of Computer Science and Engineering, IIT Patna, India,
[2]Department of Computer Science and Engineering, IIT (ISM) Dhanbad, India
[3]School of Artificial Intelligence, IIT Jodhpur, India
gopendra.99@gmail.com, sai_2001cs58@iitp.ac.in, muazama.03@gmail.com, asif@iitp.ac.in,

Abstract

Cognitive distortion research holds increasing significance as it sheds light on pervasive errors in thinking patterns, providing crucial insights into mental health challenges and fostering the development of targeted interventions and therapies. This paper delves into the complex domain of cognitive distortions which are prevalent distortions in cognitive processes often associated with mental health issues. Focusing on patient-doctor dialogues, we introduce a pioneering method for detecting and reasoning about cognitive distortions utilizing Large Language Models (LLMs). Operating within a multimodal context encompassing audio, video, and textual data, our approach underscores the critical importance of integrating diverse modalities for a comprehensive understanding of cognitive distortions. By leveraging multimodal information, including audio, video, and textual data, our method offers a nuanced perspective that enhances the accuracy and depth of cognitive distortion detection and reasoning in a zero-shot manner. Our proposed hierarchical framework adeptly tackles both detection and reasoning tasks, showcasing significant performance enhancements compared to current methodologies. Through comprehensive analysis, we elucidate the efficacy of our approach, offering promising insights into the diagnosis and understanding of cognitive distortions in multimodal settings.The code and dataset can be found here: `https://www.iitp.ac.in/~ai-nlp-ml/resources.html#ZS-CoDR`.

1 Introduction

The pervasive impact of mental health disorders (Iyortsuun et al., 2023), particularly depression and anxiety, poses significant global challenges, with substantial economic costs and profound personal suffering. The World Health Organization (WHO)[1] estimates an annual productivity loss of \$1 trillion due to these conditions. Cognitive distortions, which are inaccurate thought patterns (Dozois and Beck, 2008) contributing to negative thinking, play a crucial role in the development and exacerbation of these disorders.

While considerable research has focused on detecting cognitive distortions (Shickel et al., 2020; Singh et al., 2023; Shreevastava and Foltz, 2021), merely identifying them does not provide a comprehensive understanding of the underlying psychological processes in conversations. It is essential to elucidate the origins and thought patterns that give rise to these distortions. In Fig 1, besides the cognitive distortion label, the reasoning includes the type of negative thinking pattern and the trigger, such as the patient's statement about others' comments. The explanation of cognitive distortions (CoDs) is vital for mental health and therapeutic practices. It enhances diagnosis by providing comprehensive insights into thought patterns and triggers, allowing for contextual analysis of a patient's mental state. This understanding helps therapists design personalized interventions and aids patients in recognizing negative thinking patterns, essential in cognitive-behavioral therapy (CBT).

To equip natural language processing (NLP) systems to advance AI and automation, explanations build trust and transparency, encouraging the adoption of AI tools and ensuring decisions are ethically sound. Moreover, explanations drive research and development, leading to improved models and interventions in cognitive distortions. This paper embarks on a pioneering endeavor by curating a high-quality dataset of multimedia doctor-patient conversations annotated with cognitive distortion labels and reasoning.

Despite challenges with dataset size and human interpretation variability, we have diligently curated a reliable, labeled dataset for reasoning. This

[1]`https://www.who.int/teams/mental-health-and-substance-use/`

Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 22546–22570
November 12-16, 2024

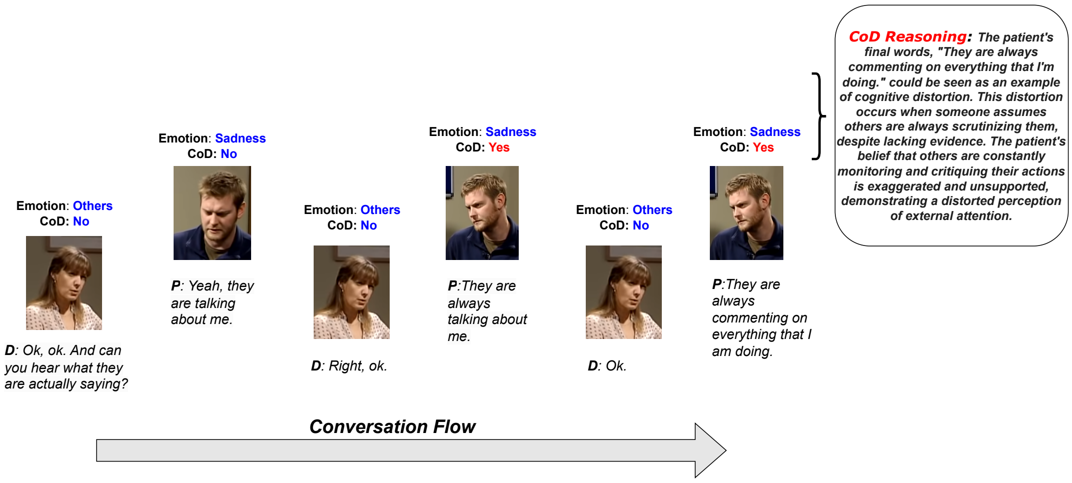


Figure 1: A conversation between Doctor and Patient, from our dataset with corresponding Emotion and Cognitive Distortion (CoD) Labels and Reasoning.

crucial contribution supports advancements in detecting and reasoning about cognitive distortions in patient-doctor dialogues. By training our model using a zero-shot approach, we aim to enable it to independently recognize subtle cues in conversations and interpret the nuanced facial expressions of patients and doctors. This method allows the model to explain cognitive distortions on its own, using contextual and interactional understanding. Our zero-shot model's improved performance over traditional methods highlights the effectiveness of this approach.

The key contributions of our work are *four-fold*: (i) We introduce a novel task i.e., Cognitive Distortion Detection and Reasoning in Conversations focusing on mental health domain; (ii) We provide a multimodal corpus, which contains doctor-patient interactions, with cognitive distortion labels and corresponding reasoning; (iii) We propose a multimodal, hierarchical framework called Zero Shot Cognitive Distortion detection and Reasoning generation (ZS-CoDR) leveraging LLMs and cross attention based modality alignment to solve both the detection and reasoning tasks; (iv) Lastly, experimental results show performance improvement compared to the baselines and provide a benchmark for our target task.

2 Related Work

Cognitive distortion is a serious mental health disorder and often is a precursor to many other disorders. The authors in Shickel et al. (2020) have compared different techniques, such as logistic regression, support vector machines, BERT, and Transformer, to detect cognitive distortion and further classify it. Although the existing work Singh et al. (2023) has incorporated multimodal patient-doctor interactions to train a multitasking framework to detect cognitive distortion, it does not address the reasoning task. Additionally, authors in Singh et al. (2023); Shreevastava and Foltz (2021) have utilized patient-doctor interactions as a dataset for their models, emphasizing their importance for training.

Our detailed literature review suggests that on mental health disorders, research focusing on reasoning generation is very limited, and in the case of cognitive distortion, there are none to the best of our knowledge. The importance of generating reasoning for the model's detection is highlighted by Gilpin et al. (2018); Ahmed et al. (2022). Moreover, the importance of incorporating multi-modal input, such as video, and audio of the patient interaction is increasing Zhang et al. (2020); Uban et al. (2022); Moreno et al. (2023); Ray et al. (2019) as it enhances the performance of the model, thereby improving the diagnosis.

Hence, by addressing these limitations, we take a step forward to solve the novel task of detecting cognitive distortion from multimodal patient-doctor interaction and generate relevant reasoning for detecting cognitive distortion. To this end, we create a new dataset and propose an effective zero-shot learning approach to solve the task.

3 Methodology

In this section, we first define the problem and then describe our proposed framework, ZS-CoDR's pipeline, and its components.Our primary objective is to classify whether a given text contains cognitive distortion or non-cognitive distortion, designated as Y, in the kth utterance $U_k = (U_{k,1}, U_{k,2}, \ldots, U_{k,n})$, where n is no. of tokens in utterance. Each utterance is associated with video V_k, and audio A_k features, all situated within the broader conversational context $H_k = ((U_1, V_1, A_1), (U_2, V_2, A_2), \ldots, (U_{k-1}, V_{k-1}, A_{k-1}))$. Furthermore, we consider the presence of emotion at the utterance level, denoted as E. Our secondary task is to generate the reasoning for detecting cognitive distortion.

Multimodal Representation: We use different encoders for each modality to represent, and later align them with the LLM's text embedding space. **Textual Encoder:** We primarily use LLAMA-7B LLM (Touvron et al., 2023) as the textual encoder. We have also shown a detailed analysis of using different LLMs.
Audio Encoder: We use the multilingual speech recognition model, WHISPER (Radford et al., 2023), to extract pertinent representations from audio data.The WHISPER model is proven effective for the English language, although it was trained for multilingual speech, as claimed by the authors in (Radford et al., 2023). Hence, we chose to work with it. Specifically, we use WHISPER-BASE to encode the audio signals.
Video Encoder: To encode video data, our strategy involves implementing a spatial-temporal contrastive learning framework, as proposed in (Qian et al., 2021). The backbone of this framework is the 3D-ResNet-50 architecture, which generates the encodings utilized in our specific task.

During the training of the 3D-ResNet-50, the spatiotemporal contrastive learning framework samples two video clips from each raw input video. A temporally consistent spatial augmentation is applied to all such sampled video clips. Since, for a given raw input video, both of its corresponding sampled clips are from the same raw input video, the RESNET3D is trained to embed them into similar vectors, using InfoNCE loss. These sampled clips are passed through the ResNet block. The resulting encodings undergo further processing in a Multi-layer Perceptron (MLP) block, culminating in a 128-dimensional vector denoted as V. The loss computation is based on the output of the MLP block (Chen et al., 2020). The core component of this learning framework is the InfoNCE (Noise Contrastive Estimation) contrastive loss proposed by (Oord et al., 2018). For a batch of size B, given feature vectors V_i and V_i' corresponding to two sampled and augmented clips from the i-th video, and a temperature parameter $\theta > 0$, the loss (L) is defined as:

$$L = \frac{1}{B} \sum_{i=1}^{B} L_i$$

where, L_i represents the loss for the i-th video:

$$L_i = -\log \frac{\exp\left(\frac{\text{sim}(V_i, V_i')}{\theta}\right)}{\sum_{k=1, k \neq i}^{2B} \exp\left(\frac{\text{sim}(V_i, V_k)}{\theta}\right)}$$

Here, $\text{sim}(V_i, V_k) = \frac{V_i \cdot V_k}{\|V_i\|_2 \cdot \|V_k\|_2}$

The advantage of this loss function lies in its capability to attract two feature vectors from the i-th video (V_i, V_i') toward each other while simultaneously repelling them from feature vectors corresponding to the other videos. Initially, the encoder is trained on our videos using this framework. The trained ResNet backbone is later employed for our main task. The contrastive loss framework was proven effective for video modality, specifically in the original paper, but no such experiments were conducted for audio modality, and LLMs have been proven effective for processing text-based modality. Hence, we applied contrastive loss only to the video modality.
Modality Alignment: Traditionally, LLMs work on textual modalities. Hence, encoding other modalities to the text embedding space for the LLMs to comprehend information from these modalities is imperative. To avoid the inherent variations in the generated representations, researchers have prominently adopted different alignment techniques to seamlessly align various modalities to the textual feature space of LLMs(Lyu et al., 2023; Alayrac et al., 2022). Hence, we employed a cross attention mechanism, which has proven effective for bridging different modality representations to textual space (Lyu et al., 2023; Alayrac et al., 2022). In our case, we align video and audio encodings with the text embedding space of LLM, similar to (Lyu et al., 2023), resulting in audio and video tokens.

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_k}}\right)\mathbf{V} \quad (1)$$

Here, $\mathbf{Q}$ represents the query matrix, $\mathbf{K}$ represents the key matrix, and $\mathbf{V}$ represents the value matrix. The function softmax is applied element-wise,

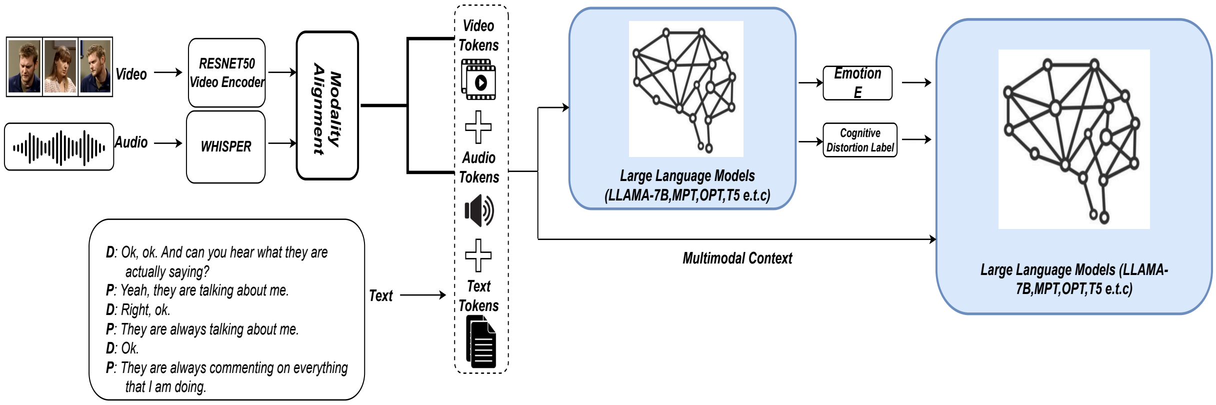


Figure 2: Architectural diagram of our proposed framework, ZS-CoDR

and d_k is the dimensionality of the key vectors and query vectors, while d_v is the dimensionality of the value vector.

Let h_v and h_a be the video and audio features representations from the respective encoders, where $h_v \in \mathbb{R}^{L_v \times d_h}$, and $h_a \in \mathbb{R}^{L_a \times d_h}$ are image, video, and audio features, respectively, and d_h is the dimension of modality-specific features. To bring them to coherent dimension space, the features are transformed using a 1-D convolutional layer, followed by a linear layer, to reduce the number of prefix tokens and align the feature size to the size of the LLMs embedding matrix. This also helps reduce the computational costs.

$$\begin{aligned} h'_v &= Linear(Conv1D(h_v)) \\ h'_a &= Linear(Conv1D(h_a)) \end{aligned} \tag{2}$$

where $h'_v \in \mathbb{R}^{L' \times d_e}$, and $h'_a \in \mathbb{R}^{L' \times d_e}$ are the transformed features with a fixed length of L' and an embedding dimension of d_e, same as the dimensionality of the embedding matrix of textual LLM. The embeddings h'_v and h'_a are then aligned with textual embedding space using attention mechanism, from eqn 1.

$$\begin{aligned} h^t_a &= \text{Attn}(h'_a, E, E) \\ h^t_v &= \text{Attn}(h'_v, E, E) \end{aligned} \tag{3}$$

where h^t_a and h^t_v are the corresponding aligned representation, and E is the embedding matrix associated with LLM. The 1D-Conv (one-dimensional convolution) is trained with an objective function designed to optimize the alignment between input features and target labels or representations. The Linear layer also requires training, as it needs to accurately map the input sequences to the aligned output sequences. This training process involves several steps: defining the objective function, which measures the alignment accuracy between the input and target sequences; training the 1D-Conv layer by adjusting the convolutional filter weights to minimize the objective function's error; and concurrently training the linear layer to ensure proper sequence alignment based on the learned weights from the 1D-Conv layer. Following this alignment procedure, the LLM can effortlessly handle representations from diverse modalities. The aligned modality representations constitutes the multimodal context and are integrated into the instruction through the process of concatenation. It can be formulated as:

$$x = [h_t : h^t_a : h^t_v : \text{Embed}(inst_t)] \tag{4}$$

where [:] denotes the concatenation operation, x signifies the multi-modal instruction, h_t represents the textual utterances, $inst_t$ corresponds to the sequence of tokens in the prompt given to LLM.

Cognitive Distortion and Emotion Prediction. We pass the multimodal context to the first LLM as shown in the Fig 2, and prompt it to predict the presence of cognitive distortion in the patient's utterance. We also prompt it to predict the emotion present in the utterance along with Cognitive Distortion detection.

Reasoning Generation After obtaining the predicted label $\hat{y}$ from the inference step along with the emotion E, the reasoning generation happens in a zero-shot manner, using a second LLM. The prompt for reasoning generation contains the following information:

1. The multimodal aligned context representation, used in the first LLM.

2. The presence of cognitive distortion and emotion in target utterance.

3. Instruction to generate reasoning for the detection of cognitive distortion, by utilizing the context provided.

The second LLM decomposes complex tasks into manageable sub-tasks (detection and reasoning), improving accuracy and performance. Initial layers detect CoDs with multimodal inputs, while subsequent layers generate detailed explanations. This approach enhances modularity, scalability, and resource utilization, aligning with human cognitive processes and improving interpretability.

4 Dataset and Experiments

4.1 Dataset.

Analyzing conversations between doctors and patients holds immense potential for training models to detect cognitive distortions. These dialogues provide a rich source of real-world language patterns used by individuals experiencing distorted thinking. By examining how patients express themselves and the doctor's responses, the model can learn to identify linguistic markers associated with specific cognitive distortions, ultimately leading to more accurate automated detection and analysis. Hence, we chose to work with the Cognitive Distortion and Emotion Cause (CoDEC) dataset used in (Singh et al., 2023). The CoDEC dataset offers 30 recordings of doctor-patient interactions, where patients exhibit various cognitive distortions like extreme thinking and overgeneralization. These conversations come in two forms: real interviews with psychiatrists and patients (20), and staged scenarios with psychiatrists and actors portraying mental health patients (10). Each interaction is linked to a YouTube video, providing synchronized video and audio data for analysis. The conversations average around 125 utterances, with sentences averaging 11.41 words.

Cognitive Distortion Annotation. In the original CoDEC dataset, each utterance is labeled with details like who spoke (doctor or patient), emotion shown at each utterance and the content type. This includes factual information ("fact"), signs of distorted thinking ("cognitive distortion"). To identify these labels, three independent annotators reviewed the utterances. The final label for each utterance was determined by a majority vote among their individual annotations.The annotators focused on identifying utterances that showed biased perspectives or irrational interpretations of real-world situations. Given the involvement of more than two annotators, a Fleiss-Kappa score(Spitzer et al., 1967) of 0.83 was calculated, indicating a high level of agreement between the annotators.

Reasoning Annotation Since, the original CoDEC dataset consists of only cognitive distortion labels but not the reasoning for the labels, we had to augment the dataset with reasoning. We employ three annotators, with a sound understanding of the phenomenon of cognitive distortion and its various forms, to provide reasoning for the cognitive distortion labels. They were asked to include parts from the context, which support the patient's distorted thinking presented in the labeled utterance, as well as use the facts from doctor's questions. Additionally, they also mentioned how the labeled utterance along with the context presents cognitive distortion in the patient. Once again, Fleiss-Kappa κ (Spitzer et al., 1967) score was used to calculate inter-annotator agreement, and we obtain a score of 0.79. Hence, using the CoDEC dataset, we augmented it with reasoning for cognitive distortion labels to create a new dataset, Cognitive Distortion Detection and Reasoning (CoDeR), to solve our task.

Challenges. Obtaining doctor-patient interactions is a huge challenge, Since doctor-patient interactions are often confidential, because of privacy and the nature of sensitivity involved in it. To our knowledge, only the CoDEC dataset was open-source and relevant to our task.

The subjective nature of annotating reasoning for cognitive distortion labels, proved to be another hurdle, with no prior cues, from the CoDEC dataset, sometimes, our annotators faced difficulty pinpointing the reason for the label. Hence, we had to discard such cases, which were around sixty. Additionally, these annotations demand a solid grasp of medical knowledge and mental health concepts.

4.2 Experimental Setup:

Owing to space limitations, we elucidate the experimental setup for **ZS-CoDR** in **Appendix** D.

Baselines: Our main goal was to evaluate a variety of techniques, especially since no existing baselines were tailored to our specific task. We focused on comparing our framework with other zero-shot learning methods to gauge its effectiveness. We use the following supervised cognitive distortion reasoning generation tasks as our baselines: MOSES (Kumar et al., 2023), KM-BART (Xing et al., 2021), One-LLM(Han et al., 2023). Zero-shot cognitive distortion reasoning generation: NMT (Lakew et al., 2018), ZSDG (Zhao and Eskenazi, 2018), and ZeroNLG (Yang et al., 2024). We assess the effectiveness of our method using the PPL and BLEU metrics against these baselines. For cognitive distortion identification, we utilize five baselines, *viz.* DialogueRCN (Hu et al., 2021), Bi-Direction RNN (Raheja and Tetreault, 2019), One-LLM (Han et al., 2023), Semantic Knowledge + Zero-Shot Classifier (Zhao and Eskenazi, 2018), and ZeroNLG (Yang et al., 2024). Further details on the baselines can be found in the Appendix (Sections B).

Evaluation Metrics: We employ various metrics for both automatic and manual evaluation purposes. For manual evaluation, we employed three distinct metrics (Singh et al., 2022), each rated on a scale from 0 to 5, focusing on Fluency, Knowledge Consistency, and Informativeness[2]. Detailed metrics explanations can be found in **Appendix** E.

5 Results and Analysis

Main Result. In Table 1, we present the results for both the tasks. The most notable observation is the consistently substantial improvement demonstrated by *ZS-CoDR* across all metrics and tasks, encompassing cognitive distortion identification (refer to Table 1) and cognitive distortion reasoning (refer to Table 1). Upon examining the table, specifically focusing on the CoDER dataset and the cognitive distortion identification task, we achieve a significant improvement of 6.17% in terms of *F1 score*(Table 1) compared to the baseline *ZeroNLG* approach.

Regarding cognitive distortion reasoning generation, we observe significant enhancements of 7.9 and 9.87 decrement(Table 1) in comparison to the baseline *ZeroNLG* approach, as indicated by the improvements in BLEU-4 and PPL scores(Table 1), respectively. Similarly, we also observe a substantial increase of 6.59 in the BERTScore. By examining a broad spectrum of architectures, including LSTM, encoder-decoder, and LLMs, we aimed to demonstrate the superior performance of our proposed framework, its alignment technique, and zero-shot learning. Additionally, the enhanced performance of our ZS-CoDR in reasoning generation underscores the potential of zero-shot learning in addressing the challenges of the

[2]Responses deemed most incorrect were assigned a score of 0, whereas the highest quality responses received a score of 5

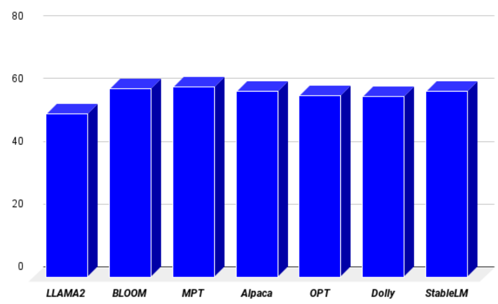


Figure 3: Comparison of different LLMs in terms of Perplexity Scores

cognitive distortion domain, which typically requires substantial knowledge. Consequently, we can confidently assert that our proposed approach, *ZS-CoDR-LLAMA7B*, when compared with the ZS-CoDR with other LLMs, as evident from Fig 3 stands out as the most effective solution for both tasks based on standard evaluation metrics.

Comparisons among different LLMS Our approach is agnostic to any specific LLM and aims to identify the most effective one among a range of options. In our study, we employed ten different LLMs: OPT, LLAMA, BLOOM, MPT, ALPACA, Vicuna, DOLLY, Stable LM, XLNET, and T5. Through rigorous experimentation, we discovered that LLAMA 7b consistently outperformed all the other LLMs in terms of various evaluation metrics. The superior performance of LLAMA 7b was evident across multiple tasks and datasets. This could be attributed to several factors, including the architecture, pre-training data, and fine-tuning strategy of LLAMA 7b, which enabled it to better capture the complexities of the cognitive distortion reasoning task. Consequently, for the purpose of our study, we selected LLAMA as the reference LLM for comparison with different baseline models. The detailed results and responses generated by different LLMs are provided in the **Appendix** to highlight the variability in perplexity and performance. Additionally, in Figure 3, we visually demonstrate that LLAMA 7B consistently yields the most favorable results among all tested LLMs, further supporting our choice for comparison.

Human Evaluation: To assess the quality of the generated reasoning by the *ZS-CoDR* model, a human evaluation was conducted using a randomly selected sample of 250 instances from the test set. Consistent with the experimental results (refer to Table 1), the outcomes of the human evaluation (see Table 2) affirm the superior performance of

Baseline	$F1^{CD}$%	Acc^{CD}%	B-4	PPL	BS	M
DialogueRCN(Hu et al., 2021)	57.00	59.98	-	-	-	-
Bi-Direction RNN(Raheja and Tetreault, 2019)	55.50	53.45	-	-		
MOSES(Kumar et al., 2023)	-	-	2.31	72.70	58.22	22.71
KM-BART(Xing et al., 2021)	-	-	6.44	68.30	56.63	24.80
One-LLM(Han et al., 2023)	66.80	77.84	14.51	64.60	59.62	31.03
SK+ZS Classifier(Zhang et al., 2019)	63.00	71.46	-	-	-	-
NMT(Lakew et al., 2018)	-	-	7.92	68.70	52.42	30.18
ZSDG(Zhao and Eskenazi, 2018)	-	-	7.81	64.10	51.88	32.82
ZeroNLG(Yang et al., 2024)	66.80	77.84	16.32	65.10	63.33	39.32
ZS-CoDR(Proposed$_{LLaMA-7B}$ +EMOCA(Daněček et al., 2022))	78.93	**86.19**	21.07	**54.73**	**70.31**	43.77
ZS-CoDR(Proposed$_{LLaMA-7B}$)	**79.57**	84.31	**22.22**	55.20	69.92	**45.23**

Table 1: Automatic evaluation results for Cognitive Distortion Detection and Reasoning. **Due to space constraint we release the score of emotion in Appendix I.2.1**. Here, B-4, M, BS, and PPL denote BLEU-4, Meteor, BERTScore, and Perplexity, respectively. Where CD: Cognitive Detection

ZS-CoDR compared to the existing baselines in generating appropriate zero-shot reasoning. It is evident that *ZS-CoDR* consistently outperforms the baselines across various manual evaluation metrics. The generated responses are not only fluent but also highly relevant to the given context, effectively encapsulating crucial information including the patient's perspective, the intended target, and the essence of cognitive distortion within the dialogue, thus providing comprehensive reasoning for cognitive distortion.

Models	Fluency	Knowledge consistency	Informativeness
MOSES	2.08	2.11	2.46
One-LLM	**2.21**	**2.29**	**2.83**
ZeroNLG	2.95	2.88	3.01
ZS-CoDR	**3.14**	**3.22**	**3.40**

Table 2: Results of human evaluation on cognitive distortion reasoning task

Case Study: In Figure 4, we present case studies illustrating zero-shot reasoning segments from the dataset within the context of the cognitive distortion reasoning task. The figure demonstrates that within the dataset, the reasoning generated by our proposed *ZS-CoDR with LLAMA-7B* framework exhibit higher accuracy, fluency, and information content compared to the baseline *ZeroNLG* approach, closely aligning with the actual ground-truth reasoning. The baseline approach tends to produce shorter reasoning, resulting in the omission of context and vital information. It is evident that our proposed approach yields improved reasoning compared to the *ZeroNLG* approach and is on par with the gold-standard reasoning provided for the given dialogue instance. Additionally, in Fig 4 we compare reasoning generated by considering all three modalities and just text modality. The reasoning generated by the multimodal model is more clear and more accurate than the plain text model.Since it is difficult to show multimodal features such as eye gaze, body language, e.t.c on the paper, the GitHub link provided in the abstract contains the YouTube links for the patient-doctor interactions in the dataset, which emphasize the importance of audio and visual cues. We also showcase different responses generated with different LLMs in the **Appendix**.

Setup	$F1^{CD}$ (%)	BS^{CR} (%)
[ZS-CoDR]	**79.57**	**69.92**
[ZS-CoDR]$_T$	75.21 (-4.36)	65.31 (-4.61)
[ZS-CoDR]$_V$	60.05 (-10.52)	60.15 (-9.77)
[ZS-CoDR]$_A$	70.88 (-8.69)	60.81 (-9.11)
[ZS-CoDR]$_{T+A}$	75.80 (-3.77)	66.43 (-3.49)
[ZS-CoDR]$_{A+V}$	73.39 (-6.18)	62.70 (-7.22)
[ZS-CoDR]$_{T+V}$	73.69 (-5.88)	64.49 (-5.43)
[ZS-CoDR]$_{-Emotion}$	77.28 (-2.29)	67.04 (-2.88)

Table 3: Results of ablated models. % fall in scores are shown in brackets. Here, CD: Cognitive Detection, CR: Cognitive Reasoning

Ablation Study: We conducted an ablation study on our proposed model(ZS-CoDR), systematically removing specific components such as multimodal features and emotions. Table 3 signifies the ablation study by including different combinations of modality, instead of all 3 together. Similarly, the last row in Table 3 refers to the removal of emotional components from the proposed architec-

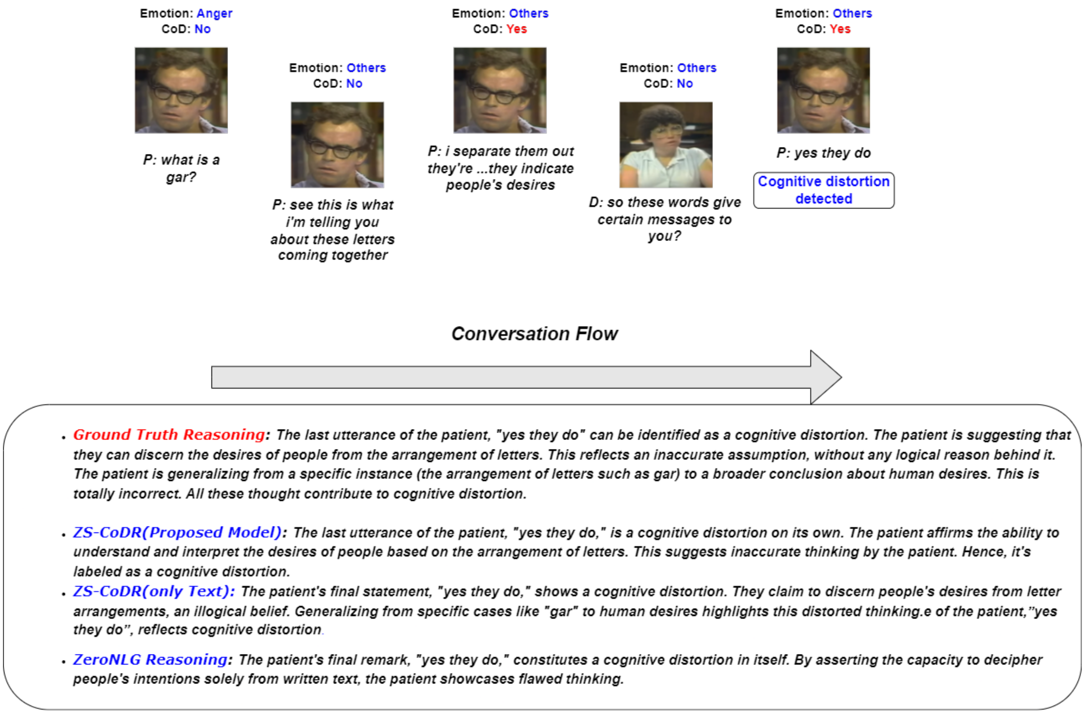


Figure 4: Comparisons among ground truth reasoning and reasoning generated by our model ZS-CoDR and zero-shot baseline ZERONLG. Additionally, we also generate resasoning using ZS-CoDR with only Tect modality. ZS-CoDR's(multimodal) response is better aligned with the ground truth as it mentions the patient's remark on the arrangement of letters and links it with cognitive distortion.ZS-CoDR's(only Text) response falls short in comparison to the multimodal in terms of coherence with ground truth and clarity.While ZeroNLG's response is more generic and not very informative.

ture. The results presented in this table emphasize the pivotal role of each component. Various combinations were examined, including multimodal features with emotions, only text input, and others. The observed decrease in performance metrics upon component removal underscores the significance of each component's contribution to the overall model performance. From Table 3, it is evident that between video and audio, video has more impact. But overall, combining all three modalities has superior performance than other combinations, as evident from higher evaluation metric scores in first-row and last-row models. Hence, we specifically incorporated multiple modalities because relying on a single modality is insufficient for understanding the complexity of a patient's thoughts and behaviors. By considering audio, video, and text data, our model gains a more comprehensive understanding of the patient's state, allowing for more accurate and insightful responses. The emotion component also helps in improving the performance in both tasks, as evident in the decrease in performance, by removing the emotion component in last row of Table 3

6 Conclusion

In our paper, we have addressed a very vital task of zero-shot response generation for cognitive distortion, essential for comprehending altered behavior and its underlying reasons. Large Language Model (LLM) conditioned on predicted labels and multimodal input data, including audio, video, and text. Utilizing LLM's architecture, our model processes multi-modal data and generates coherent, contextually relevant responses without task-specific training. Experimental results validate our approach's effectiveness, indicating its potential to offer valuable insights into cognitive distortion across diverse domains, fostering better understanding and facilitating nuanced analysis. Our current dataset contains around 743 Cognitive Distortion utterances. Most of these utterances are from patients suffering from Psychosis or Paranoid Schizophrenia, and a lesser no. of patients suffering from depression and personality disorders. Future works can further increase the utterance to capture more dimensions of cognitive distortion and conduct analysis on the sub-classes of cognitive distortion.

7 Limitations

In addition to the aforementioned points, it's crucial to acknowledge that the nature of patient-doctor dialogues is unique, often involving nuanced communication dynamics and specialized terminology. This specificity could potentially limit the effectiveness of the proposed method when applied to other types of conversations, such as those in legal or educational settings.

Moreover, the ethical considerations surrounding the use of multimodal data extend beyond mere technical implementation. In sensitive domains like mental health, where confidentiality and trust are paramount, the responsible handling of data becomes even more critical. Issues such as the inadvertent disclosure of sensitive information or the potential for algorithmic biases to exacerbate existing disparities in healthcare access and treatment outcomes must be thoroughly addressed.

Furthermore, while the study may demonstrate promising results within its controlled environment, the real-world variability of conversational data poses challenges to generalization. Factors such as diverse linguistic styles, cultural nuances, and contextual cues can significantly impact the performance of any automated system. Therefore, ongoing validation efforts across a wide range of datasets and conversational contexts are essential to ensure the reliability and effectiveness of the proposed method in diverse real-world scenarios.

8 Ethical Considerations

The rigorous evaluation and review conducted by our Institutional Review Board (IRB) ensure that the study adheres to strict ethical standards and safeguards the rights and well-being of all involved parties. It's important to emphasize that the primary objective of this research is to enhance the capabilities of medical professionals in diagnosing and addressing medical health issues, ultimately leading to improved patient care and overall human well-being. By leveraging innovative technologies and methodologies, the study aims to empower healthcare providers with valuable insights and tools to enhance medical practice and outcomes.

Regarding the utilization of YouTube videos in the dataset, it's worth noting that these videos are sourced responsibly and ethically. They are freely available online without any copyright restrictions, and their usage is solely for research and educational purposes. Furthermore, the dissemination of these videos through various channels serves the overarching goal of advancing scientific knowledge and fostering educational initiatives within the medical community. This transparent approach ensures compliance with legal and ethical guidelines while promoting the exchange of information and collaboration in the pursuit of scientific advancement.

References

Usman Ahmed, Rutvij H Jhaveri, Gautam Srivastava, and Jerry Chun-Wei Lin. 2022. Explainable deep attention active learning for sentimental analytics of mental disorder. *Transactions on Asian and Low-Resource Language Information Processing*.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. 2022. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736.

Megha Chakraborty, SM Tonmoy, SM Zaman, Krish Sharma, Niyar R Barman, Chandan Gupta, Shreya Gautam, Tanay Kumar, Vinija Jain, Aman Chadha, et al. 2023. Counter turing test ctˆ 2: Ai-generated text detection is not as easy as you may think–introducing ai detectability index. *arXiv preprint arXiv:2310.05030*.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR.

Radek Daněček, Michael J Black, and Timo Bolkart. 2022. Emoca: Emotion driven monocular face capture and animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20311–20322.

David JA Dozois and Aaron T Beck. 2008. Cognitive schemas, beliefs and assumptions. *Risk factors in depression*, pages 119–143.

Leilani H Gilpin, David Bau, Ben Z Yuan, Ayesha Bajwa, Michael Specter, and Lalana Kagal. 2018. Explaining explanations: An overview of interpretability of machine learning. In *2018 IEEE 5th International Conference on data science and advanced analytics (DSAA)*, pages 80–89. IEEE.

Jiaming Han, Kaixiong Gong, Yiyuan Zhang, Jiaqi Wang, Kaipeng Zhang, Dahua Lin, Yu Qiao, Peng Gao, and Xiangyu Yue. 2023. Onellm: One framework to align all modalities with language. *arXiv preprint arXiv:2312.03700*.

Dou Hu, Lingwei Wei, and Xiaoyong Huai. 2021. Dialoguecrn: Contextual reasoning networks for emotion recognition in conversations. *arXiv preprint arXiv:2106.01978*.

Ngumimi Karen Iyortsuun, Soo-Hyung Kim, Min Jhon, Hyung-Jeong Yang, and Sudarshan Pant. 2023. A review of machine learning and deep learning approaches on mental health diagnosis. In *Healthcare*, volume 11, page 285. MDPI.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Shivani Kumar, Ishani Mondal, Md Shad Akhtar, and Tanmoy Chakraborty. 2023. Explaining (sarcastic) utterances to enhance affect understanding in multimodal dialogues. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 12986–12994.

Surafel M Lakew, Quintino F Lotito, Matteo Negri, Marco Turchi, and Marcello Federico. 2018. Improving zero-shot translation of low-resource languages. *arXiv preprint arXiv:1811.01389*.

Chenyang Lyu, Minghao Wu, Longyue Wang, Xinting Huang, Bingshuai Liu, Zefeng Du, Shuming Shi, and Zhaopeng Tu. 2023. Macaw-llm: Multi-modal language modeling with image, audio, video, and text integration. *arXiv preprint arXiv:2306.09093*.

Felipe Moreno, Sharifa Alghowinem, Hae Won Park, and Cynthia Breazeal. 2023. Expresso-ai: An explainable video-based deep learning models for depression diagnosis. In *2023 11th International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 1–8. IEEE.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

Rui Qian, Tianjian Meng, Boqing Gong, Ming-Hsuan Yang, Huisheng Wang, Serge Belongie, and Yin Cui. 2021. Spatiotemporal contrastive video representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6964–6974.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *International Conference on Machine Learning*, pages 28492–28518. PMLR.

Vipul Raheja and Joel Tetreault. 2019. Dialogue act classification with context-aware self-attention. *arXiv preprint arXiv:1904.02594*.

Anupama Ray, Siddharth Kumar, Rutvik Reddy, Prerana Mukherjee, and Ritu Garg. 2019. Multi-level attention network using text, audio and video for depression prediction. In *Proceedings of the 9th international on audio/visual emotion challenge and workshop*, pages 81–88.

Benjamin Shickel, Scott Siegel, Martin Heesacker, Sherry Benton, and Parisa Rashidi. 2020. Automatic detection and classification of cognitive distortions in mental health text. In *2020 IEEE 20th International Conference on Bioinformatics and Bioengineering (BIBE)*, pages 275–280. IEEE.

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.

Gopendra Vikram Singh, Mauajama Firdaus, Shruti Mishra, Asif Ekbal, et al. 2022. Knowing what to say: Towards knowledge grounded code-mixed response generation for open-domain conversations. *Knowledge-Based Systems*, 249:108900.

Gopendra Vikram Singh, Soumitra Ghosh, Asif Ekbal, and Pushpak Bhattacharyya. 2023. Decode: Detection of cognitive distortion and emotion cause extraction in clinical conversations. In *European Conference on Information Retrieval*, pages 156–171. Springer.

Robert L Spitzer, Jacob Cohen, Joseph L Fleiss, and Jean Endicott. 1967. Quantification of agreement in psychiatric diagnosis: A new approach. *Archives of General Psychiatry*, 17(1):83–87.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Ana-Sabina Uban, Berta Chulvi, and Paolo Rosso. 2022. Explainability of depression detection on social media: From deep learning models to psychological interpretations and multimodality. In *Early Detection of Mental Health Disorders by Social Media Monitoring: The First Five Years of the eRisk Project*, pages 289–320. Springer.

Yiran Xing, Zai Shi, Zhao Meng, Gerhard Lakemeyer, Yunpu Ma, and Roger Wattenhofer. 2021. Kmbart: Knowledge enhanced multimodal bart for visual commonsense generation. *arXiv preprint arXiv:2101.00419*.

Bang Yang, Fenglin Liu, Yuexian Zou, Xian Wu, Yaowei Wang, and David A Clifton. 2024. Zeronlg: Aligning and autoencoding domains for zero-shot multimodal and multilingual natural language generation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Jingqing Zhang, Piyawat Lertvittayakumjorn, and Yike Guo. 2019. Integrating semantic knowledge to tackle zero-shot text classification. *arXiv preprint arXiv:1903.12626*.

Ziheng Zhang, Weizhe Lin, Mingyu Liu, and Marwa Mahmoud. 2020. Multimodal deep learning framework for mental disorder recognition. In *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)*, pages 344–350. IEEE.

Tiancheng Zhao and Maxine Eskenazi. 2018. Zero-shot dialog generation with cross-domain latent actions. *arXiv preprint arXiv:1805.04803*.

Frequently asked questions

- **Why is explanation task important for cognitive distortion?**

  Response: explanation in the context of cognitive distortions is both an attractive and challenging task that demands further exploration. The explanation of cognitive distortions (CoDs) is crucial for several reasons, particularly in mental health and therapeutic practices. Providing explanations enhances understanding and diagnosis by offering comprehensive insights into underlying thought patterns and triggers, which is essential for accurate diagnosis. It also allows for contextual analysis, giving clinicians a deeper understanding of the patient's mental state and contributing factors. Therapeutically, understanding the reasoning behind CoDs enables therapists to design targeted and effective interventions, leading to more personalized treatment plans. It also helps patients become aware of their negative thinking patterns, a critical step in cognitive-behavioral therapy (CBT) where they learn to identify and challenge these thoughts. Enhanced communication is another benefit, as detailed explanations aid healthcare providers in clearly conveying the nature and impact of distorted thoughts to patients, and support thorough documentation and reporting for progress tracking and case reviews. In the realm of AI and automation, explanations build trust in technology, making clinicians and patients more likely to adopt AI tools if they understand the rationale behind outputs. Explanations also contribute to model transparency, ensuring decisions are based on sound reasoning, which is crucial for ethical considerations and regulatory compliance. Finally, in research and development, explanations drive further research by providing insights into cognitive distortions, helping develop more sophisticated models and interventions, and enabling better benchmarking and evaluation of different approaches. This can lead to significant improvements in existing methods and the development of new techniques.

- **Why do we need a second LLM? Why not generate the reasonings together in the first LLM?**

  Response: We use a hierarchical model as it is necessary for our particular task. In the first layer, we aim to detect cognitive distortions (CoDs), and with the help of subsequent layers, if the utterance contains a cognitive distortion, only then will it explain the CoD. The hierarchical model was used in this context for several compelling reasons. Firstly, it addresses structured complexity by allowing the decomposition of complex tasks into manageable sub-tasks, where different layers handle detection, contextual analysis, and reasoning generation, enhancing overall performance. Enhanced accuracy and performance are achieved as initial layers focus on detecting CoDs using multimodal data inputs (audio, video, text), leveraging the strengths of each modality. Subsequent layers are dedicated to generating explanations, providing detailed and contextually relevant outputs. The model's modularity and flexibility allow independent development and training of different modules, making fine-tuning easier and enhancing scalability. The approach aligns with cognitive processes, mimicking human cognition where higher-order reasoning builds upon basic functions, leading to more natural outputs. Efficient resource utilization is facilitated by focused resource allocation to different layers, reducing computational load and improving processing speed. The model also excels in handling multimodal data by integrating inputs into a common representation space, which is then used for complex tasks like reasoning. Lastly, enhanced interpretability is achieved through layer-wise analysis, helping to understand how different input data types contribute to final outputs, thereby increasing the transparency and trustworthiness of the model. In summary, the hierarchical model was chosen for its structured and efficient handling of complex tasks, enhanced detection accuracy, detailed explanations, modular development, and effective integration of multimodal data, all crucial for detecting and explaining cognitive distortions.

- **How do zero-shot cognitive reasoning models handle tasks or topics that are not explicitly provided in the prompt?**

  Response: Zero-shot cognitive reasoning models leverage their pre-trained knowledge to generalize reasonings to unseen tasks or topics. They use their understanding of language and concepts to generate reasonings based on the input they receive, even if it's outside their training data.

- **Are there any strategies for optimizing the performance of zero-shot cognitive reasoning models?**

  Response: Strategies for optimizing the performance of zero-shot cognitive reasoning models may include fine-tuning on specific reasoning tasks or domains, adjusting model hyperparameters, or incorporating additional context or information into the input.

- **Can zero-shot cognitive reasoning models understand and generate reasonings in multiple languages?**

  Response: Yes, zero-shot cognitive reasoning models can be trained on multilingual data and are capable of generating reasonings in multiple languages based on their pre-trained understanding of language and concepts.

- **How do zero-shot cognitive reasonings models deal with ambiguity or complex prompts?**

  Response: Zero-shot cognitive reasonings models use their contextual understanding and reasoning abilities to interpret ambiguous or complex prompts and generate reasonings that best match the input they receive. They may rely on probabilistic reasoning and language understanding techniques to address ambiguity.

- **What are some real-world applications of zero-shot cognitive reasoning?**

  Response: Real-world applications of zero-shot cognitive reasoning include natural language understanding systems, chatbots, question answering systems, and explainable AI applications where generating human-like reasonings is important for user interaction and transparency.

- **How can zero-shot cognitive reasoning models be fine-tuned or adapted for specific tasks or domains?**

  Response: Zero-shot cognitive reasoning models can be fine-tuned or adapted for specific tasks or domains by providing task-specific training data or prompts during the fine-tuning process. This helps the model learn to generate more accurate and contextually relevant reasonings for the target task or domain.

- **Why we chooses few older baselines also?**

  Response: Including older baselines in comparative studies serves multiple purposes. Firstly, they act as established benchmarks, representing well-established methods or models in the field, against which researchers can compare their new approaches to demonstrate improvements or advancements. Secondly, the inclusion of older baselines ensures continuity of evaluation, allowing for direct comparison with prior research and maintaining consistency in the evaluation process. Thirdly, older baselines may still perform reasonably well on certain tasks or datasets, providing a reference point for understanding the performance of newer approaches relative to established methods. Additionally, the inclusion of older baselines offers valuable historical context, aiding in understanding the progression of research in a particular area and tracing the evolution of methods and models over time. Lastly, it enables comparison across different time periods, allowing researchers to assess how the performance of new approaches compares not only with the latest methods but also with those developed at various points in time, thus providing insights into the pace of progress in the field.

- **If we use AI assistance?**

  Response Certainly, AI assistance was utilized for few paraphrasing.

A Appendix

We delve into the implementation particulars and provide comprehensive details regarding the con-

Figure 5: World cloud for annotated reasonings in CoDER dataset

Figure 6: Word Cloud for utterances in the CoDER dataset

sidered baselines and the metrics used for human evaluation. Furthermore, we conduct a detailed qualitative analysis, offering vivid comparisons between the predictions made by our model and those of the top-performing baselines.

B Baselines

We categorize the baselines into two distinct groups: those designed for the detection of cognitive distortion and those intended for generating reasonings of cognitive distortion in a zero-shot manner. The description of each baseline is provided below, organized according to their respective tasks.

1. **Cognitive Distortion Detection Task:**
    - We compare our proposed approach with leading baselines for the cognitive distortion detection task.
    - We begin by comparing our method with a range of techniques, starting from simpler methods to more complex ones:
        - **LSTM-based DialogueRCN** (Hu et al., 2021): This method relies on Recurrent Neural Networks (RNNs), specifically Long Short-Term Memory (LSTM) units, for cognitive distortion detection. LSTMs are a type of RNNs capable of capturing long-range dependencies in sequential data, making them suitable for analyzing conversational data.
        - **Bi-directional RNN** (Raheja and Tetreault, 2019): Another straightforward approach that utilizes bi-directional RNNs. Bi-directional RNNs process input sequences in both forward and backward directions, allowing them to capture context from both past and future states, which can be beneficial for understanding dialogue context.
        - **Semantic Knowledge Integrated Two-Phase Zero-Shot Classifier** (Zhang et al., 2019): This method integrates semantic knowledge into a two-phase zero-shot classification setup. It leverages external semantic knowledge sources to improve the model's understanding of conversational data, enabling better classification of cognitive distortions.
        - **Standard Encoder-Decoder Based Zero-Shot Classifier** (Yang et al., 2024): This technique employs an encoder-decoder architecture for zero-shot classification. It encodes input dialogues into a fixed-dimensional representation and decodes them into output labels, allowing the model to classify cognitive distortions without prior training on specific labels.
        - **LLM-based Technique** (Han et al., 2023): This method utilizes a powerful Large Language Model (LLM) for cognitive distortion detection. LLMs, such as GPT (Generative Pre-trained Transformer) models, are pre-trained on large amounts of text

data and fine-tuned for specific tasks, making them effective at capturing complex patterns in dialogue data.

2. **Reasoning Generation Task:**
   - Similarly, for the reasoning generation task, we compare with a mix of supervised and zero-shot settings.
   - We compare with various baseline techniques, each employing different methodologies:
     - **MOSES** (Kumar et al., 2023): MOSES utilizes a Multimodal context-aware attention technique coupled with BART (Bidirectional and Auto-Regressive Transformers) encoder-decoder architecture for reasoning generation. It leverages both textual and visual information to generate context-aware reasonings, enhancing the model's understanding of complex concepts.
     - **KM-BART** (Xing et al., 2021): KM-BART leverages knowledge from COMET and utilizes BART backbone for reasoning generation. By incorporating external knowledge from COMET (Commonsense Knowledge Enhanced Pre-training for Knowledge Graph Completion), KM-BART enhances its reasoning capabilities, leading to more comprehensive reasonings.
     - **One-LLM Technique**: This approach uses a single Large Language Model (LLM) as a baseline for reasoning generation by utilizing it in a hierarchical fashion. The model generates reasonings based on its learned representations and contextual understanding.
     - **Baselines for Zero-Shot Reasoning Generation**:
       * **NMT** (Lakew et al., 2018): NMT utilizes a training-inference-training cycle to generate reasoning in a zero-shot setting. It trains the model on a combination of labeled and unlabeled data and iteratively refines the model's parameters to improve reasoning generation.
       * **ZSDG** (Zhao and Eskenazi, 2018): ZSDG utilizes domain description and context input to generate reasonings using an action-matching training technique. It matches the generated reasonings with predefined actions, ensuring that the reasonings are contextually relevant and actionable.
       * **ZeroNLG**: ZeroNLG is used as a baseline for the reasoning task due to its encoder-decoder framework. It encodes input dialogues and decodes them into reasoning, similar to other encoder-decoder models, making it a suitable baseline for comparison.

C Data

The dataset, CoDeR is split into training sets, validation sets, and test sets. Each split includes text, audio, and video modalities for every dialogue. Table 4 contains the dataset statistics.

C.1 Word Cloud

Cognitive distortions are patterns of thinking that are irrational or inaccurate, often leading to negative emotions and behaviors. To analyze the language associated with cognitive distortions, we utilize word clouds to visually represent the frequency of words in both "Cognitive Distortion" and "Cognitive Distortion Reasoning" contexts. In these word clouds, the size of each term corresponds to its frequency in user descriptions, providing a visual representation of the most common words used in each context. Figures 5 and 6 depict the word clouds generated from the most frequent words for both cognitive distortion scenarios. This visual analysis allows for a better understanding of the language patterns associated with cognitive distortions and their reasoning.

C.2 Annotation Evaluation

Fleiss' Kappa for the generation annotations, with K-different annotators, was calculated through a systematic process. First, we constructed a rating matrix where each row represented an item and each column indicated the number of annotators who assigned that item to each possible category. Next, we calculated the proportion of all annotations that fell into each category across all items.

For each item, we then computed the agreement among the K-annotators, determining how consistently they assigned the same category using the formula $P_i = \frac{1}{m(m-1)}\left(\sum_{j=1}^{k} N_{ij}^2 - m\right)$, where m is the number of annotators, k is the number of categories, and N_{ij} is the number of annotators who assigned the i-th item to the j-th category. We averaged these agreement values across all items to obtain the mean observed agreement $\bar{P}$. We then calculated the expected agreement assuming random category assignment according to overall category proportions, using the formula $P_e = \sum_{j=1}^{k} p_j^2$, where p_j is the proportion of annotations in category j. Finally, we computed Fleiss' Kappa with the formula $\kappa = \frac{\bar{P}-P_e}{1-P_e}$, which reflects inter-rater reliability, adjusting for chance agreement. This comprehensive approach ensures the Kappa value accurately represents the consistency among the K-annotators in assigning categories while accounting for chance agreement.

D Experiment Setup

ZS-CoDR is developed using PyTorch [3], a Python-based deep learning package. We utilize the different LLM models imported from the Hugging Face Transformers [4] package for our experiments. All experiments are conducted on an NVIDIA Tesla V100-PCIE GPU. Pre-training is carried out for 7 epochs, followed by fine-tuning for 4 epochs. Optimization is performed using the Adam optimizer (Kingma and Ba, 2015), with learning rates set to 0.0003 and 0.005, and exponential decay rates (beta) of (0.9,0.999) for both tasks.

E Evaluation Metrics

We employ both automatic and manual evaluation metrics for assessing our proposed framework. For automatic evaluation, metrics such as Accuracy and F1 score are utilized. We calculate the F1 score by analyzing the context and then applying it to the specific utterance. Since the CoD label is present in that particular utterance, and our system predicts the CoD label based on that utterance alone after reading the context, the F1 score is determined accordingly. However, as correctly pointed out, multiple utterances come from the same patient/interview. To minimise the effect of users, we also calculate the F1 score for each patient and then average these scores. The final F1 score, after averaging across patients, is 71.54. When it comes to cognitive reasoning generation, we rely on standard generative task metrics such as Perplexity, BLEU-4, and METEOR. Additionally, we incorporate the multilingual version of BERTScore to gauge semantic similarity.

E.1 Automatic Evaluation-based Metrics

- **BLEU-4 (Bilingual Evaluation Understudy-4):** BLEU-4 is a standard metric for evaluating the quality of machine-translated text. It measures the n-gram overlap between the generated text and reference translations, with higher scores indicating better agreement.

- **METEOR:** METEOR (Metric for Evaluation of Translation with Explicit Ordering) is an automatic evaluation metric for machine translation. It considers precision, recall, and alignment between the generated and reference translations, incorporating synonymy and stemmed matches for a nuanced assessment of translation quality.

- **BERTScore:** BERTScore evaluates the quality of text generated by neural language models, such as BERT. It computes similarity between embeddings of generated and reference text segments using contextual embeddings from BERT, capturing semantic similarity more effectively than traditional n-gram overlap metrics.

- **Perplexity:** Perplexity is a metric commonly used to evaluate the performance of language models. It measures how well a language model predicts a given sequence of words. A lower perplexity score indicates better performance, suggesting the language model is better at predicting the next word in a sequence.

E.2 Human Evaluation-based Metrics

- *Fluency*: This determines whether or not the extracted span is fluent and natural. Natural and regular answers get a score of 5, whereas inarticulate ones receive a 0.

- *Knowledge consistency*: This metric determines how well the generated reasoning reflects the appropriate knowledge, i.e., cognitive distortion domain in our case. A score of 0 represents that the reasoning generated

[3] https://pytorch.org/

[4] https://huggingface.co/docs/transformers/index

Table 4

Attribute	Count
CoD	743
ReCoD	410
One Cause	410
Two Causes	179
Three Causes	36

(a) Emotion and Cause distribution.

Class	Count	# Causes
Anger	184	One: 101; Two: 42; Three: 10
Disgust	77	One: 49; Two: 22; Three: 2
Fear	169	One: 96; Two: 32; Three: 6
Joy	128	One: 28; Two: 7; Three: 2
Sadness	503	One: 198; Two: 80; Three: 10
Surprise	176	One: 78; Two: 24; Three: 2
Neutral	2516	No causal spans exists

Table 5: Frequency of utterances over various attributes. CoD: Cognitive Distortion; ReCoD: Response to CoD (Singh et al., 2023)

does not reflect that it belongs to the cognitive distortion domain, and subsequent scores from 1 to 5 indicate increasing consistency with the cognitive distortion domain, with 5 implying that it reflects all aspects of cognitive distortion.

- *Informativeness*: This metric captures how well the reasoning generated is able to use the context provided to accurately calculate the indicators for cognitive distortion in a patient's utterance. A score of 0 represents that the reasoning generated is uninformative and doesn't convince the user regarding the presence of cognitive distortion, while scores starting from 1 to 5 indicate that the reasoning is able to understand and capture relevant phrases from dialogue context that trigger the presence of cognitive distortion, in an increasing fashion.

F Varying Context Length.

By changing context sizes(ψ), we examine the role that context plays in the Cognitive Distortion Detection and Reasoning generation task. The following context lengths were trained for by ZS-CoDR: 1, 3, 5, 7, 9, 10. The results are represented in Figure 7. Here, 1 means there is no context, and the model merely receives the target utterance as input. We observe a steady improvement in performance as the number of previous utterances increases. When the ψ is set to 5, we get the best results. More context does not provide useful information, resulting in model confusion and poor performance.

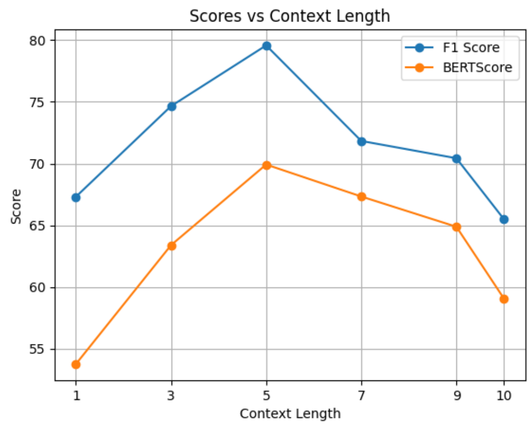


Figure 7: Graphical depiction of results of **ZS-CoDR** on varying context length.

G Case Study

We aim to illustrate the diverse responses generated by various Large Language Models (LLMs) using different figures. In Figures 8 through 15, we present the reasonings generated by ZS-CoDE with different LLMs across various conversations.

Each figure showcases a specific conversation scenario, with the reasoning provided by ZS-CoDE alongside the responses generated by different LLMs. By visualizing these responses, we gain insights into the variability and nuances in the way each LLM interprets and responds to the given conversation context.

Figure 8 to Figure 15 serve as illustrative examples of the diverse range of responses produced by different LLMs when presented with similar conversational prompts. These figures highlight the importance of considering the role of LLMs in shaping the nature and quality of generated responses, thereby providing valuable insights into the performance and capabilities of each model.

Through the analysis of these figures, we can discern patterns, trends, and discrepancies in the responses generated by different LLMs. This comparative analysis facilitates a deeper understanding of the strengths and limitations of each model and informs future research directions aimed at improving response generation in conversational AI systems.

In summary, the visual representation of reasonings generated by ZS-CoDE with different LLMs offers a comprehensive overview of the variability in response quality across different conversation contexts, thereby enriching our understanding of LLM behavior and performance in conversational settings.

H Perplexity Estimation

We compared the reasonings generated by different LLMs with the Human annotated reasonings, based on their perplexities. Taking inspiration from (Chakraborty et al., 2023), we generated 1000 bootstrapping samples, each containing 264 dialogues(reason explained soon). We plotted[5] the histogram plots of average perplexity from each bootstrapped sample in Tables 9, 10

H.1 Generating Human-Text Perplexity

- To calculate perplexity for human-annotated reasonings, we split our dataset of 660 dialogues into train and test sets in a 60:40 ratio.

- We computed the probabilities of words from the train set and utilized these probabilities to calculate perplexities for word sequences in the test set.

- The perplexity of a word sequence is computed using the formula:

$$\textbf{\textit{Perplexity}} = e^{-\frac{1}{N}\sum_{i=1}^{N}\log_e(p(w_i))}$$

  where N represents the length of the word sequence, and $p(w_i)$ denotes the probability of the individual word w_i.

- In the event of encountering out-of-vocabulary words in the test set, we assigned a small default probability.

- During the bootstrap method, we employed the test set of size 264 (40% of 660) as the original dataset to generate bootstrap samples of the same size.

I Comparison between Human and LLMs

Tables 9, 10 illustrates the comparison between text generated by various LLMs and human-generated text in terms of perplexity. Remarkably, the perplexity graph exhibits a striking similarity between ChatGPT 3.5 and LLAMA-7B, as evidenced by their nearly identical profiles. However, when comparing these results with those obtained from other LLMs (as shown in Table 10), a noticeable disparity emerges.

This observation underscores a significant finding: ChatGPT and LLAMA, even in a zero-shot manner where they possess only a rudimentary understanding of cognitive distortion, produce responses that closely resemble those generated by humans. This alignment in response quality highlights the remarkable capability of these models to capture the essence of cognitive distortion, despite lacking in-depth domain-specific knowledge.

However, it is noteworthy that LLAMA, particularly when lacking multimodal input, experiences shortcomings in certain cases. This limitation becomes apparent when considering the crucial role played by non-verbal cues, such as facial expressions of patients and body language of doctors, in understanding cognitive distortion. In such instances, the absence of multimodal information impedes LLAMA's ability to fully grasp the nuances of cognitive distortion, leading to suboptimal performance.

In summary, while ChatGPT and LLAMA demonstrate promising capabilities in generating responses akin to human-generated text, the integration of multimodal information emerges as a critical factor in enhancing model performance, particularly in contexts where non-verbal cues play a significant role.

I.1 Generated Zero-shot Reasoning by Various LLMs

In Table 7, we present various reasoning generated by different LLMs, shedding light on their respective performances. Notably, our analysis reveals that the lack of zero-shot capabilities adversely impacts the quality of responses across all LLMs. Each LLM tends to generate responses in line with

[5] https://colab.research.google.com/drive/1CBzGhc9Pj4fjmRDCqXSq1_CoL9a8erfz?usp=sharing

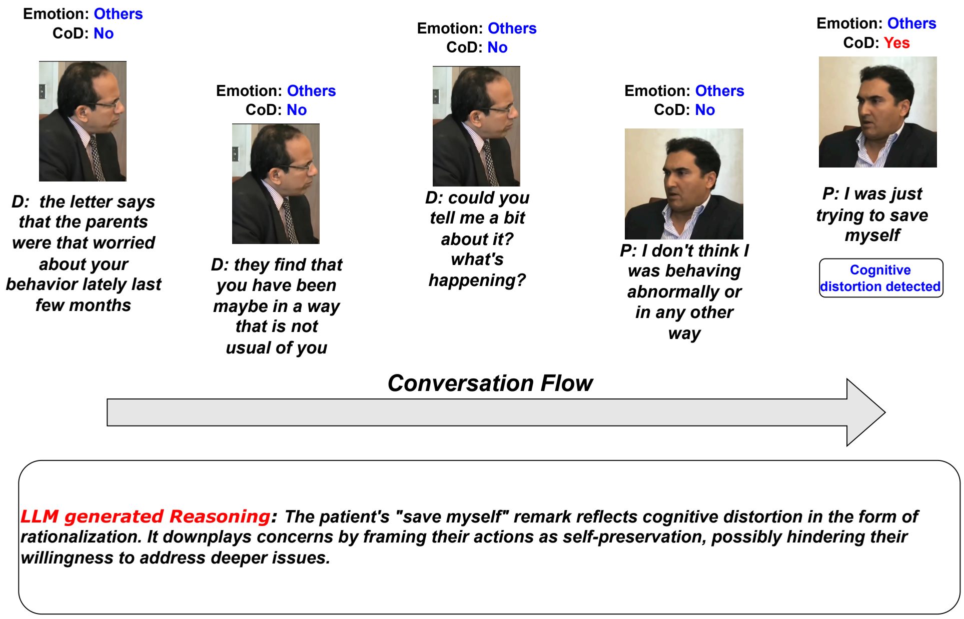


Figure 8: An example of reasoning generated by ZS-CoDR, with OPT LLM. The response, although small, is very clear in its reasoning, and highlights which phrase(here "save myself") supports the presence of cognitive distortion.

its training data, reflecting the limitations of their pre-existing knowledge.

For instance, the **Alpaca** model, trained specifically for generating creative responses, consistently produces imaginative and unconventional reasoning. This behavior aligns with its training objective and highlights its proficiency in delivering creative outputs. However, despite this specialization, the reliance on pre-existing training data constrains Alpaca's ability to adapt to novel contexts or tasks, resulting in a lack of versatility.

This observation underscores the importance of zero-shot learning, which empowers models to generalize across diverse domains and tasks without the need for explicit training. Models equipped with zero-shot capabilities exhibit enhanced flexibility and adaptability, enabling them to generate responses that align more closely with the specific requirements of a given task or context.

In summary, while specialized models like **Alpaca** excel in certain domains due to their tailored training objectives, their performance is inherently limited by their pre-existing knowledge. The integration of zero-shot learning capabilities is crucial for overcoming these limitations and enabling LLMs to generate responses that are more contextually relevant and adaptable across diverse scenarios.

I.2 Generated Zero-shot Reasoning by ChatGPT

In Table 8, we present a pilot study conducted using ChatGPT[6] to assess the effectiveness of our proposed framework. The table showcases a selection of sample predictions generated by ChatGPT in response to a given prompt.

For this pilot study, we provided ChatGPT with the following prompt:

Cognitive distortions are inaccurate thought patterns, beliefs, or perceptions that contribute to negative thinking, which subsequently elevates the chances of several mental illnesses. In the conversation between a doctor (D) and his patient (P), given below, the last utterance of the patient is labeled as cognitive distortion. Provide reasoning as to why it has been labeled as cognitive distortion.

The provided prompt sets the context for the conversation between the doctor (D) and the patient (P), highlighting the concept of cognitive distortions and their impact on negative thinking and mental health. The last utterance of the patient is designated as a cognitive distortion, and ChatGPT is tasked with generating reasoning to support this label.

In the conversation snippet provided, the patient expresses a lack of motivation or will to engage in activities, stating, "I don't have the will to do anything." This statement reflects a negative and

[6] https://chat.openai.com/

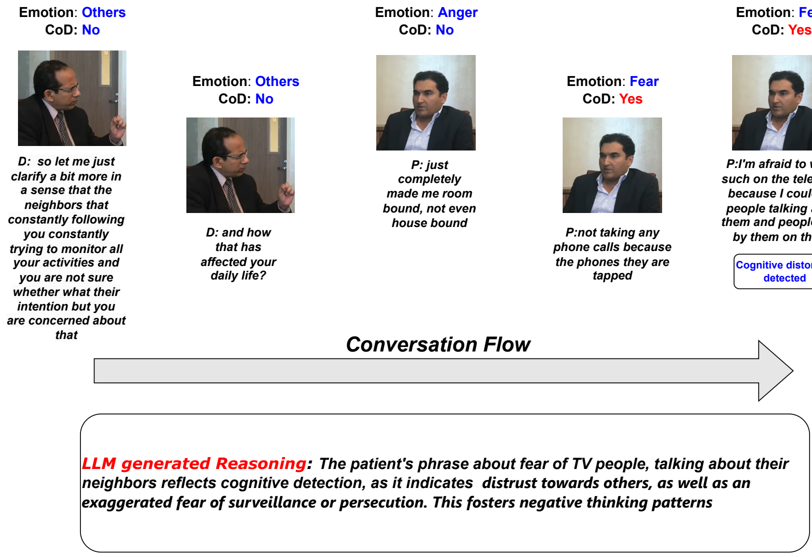


Figure 9: An example of reasoning generated by ZS-CoDR, with MPT LLM.The response, just like OPT, is crisp and clear and captures relevant phrases from context to generate the reasoning, like in this case, it mentions "patient's phrase about fear of TV people".

defeatist attitude, indicative of distorted thinking patterns associated with cognitive distortions. ChatGPT is expected to analyze this utterance in the context of cognitive distortions and provide reasoning to elucidate why it qualifies as such.

The responses generated by ChatGPT are evaluated based on their relevance, coherence, and alignment with the concept of cognitive distortions. This pilot study serves as a preliminary assessment of ChatGPT's capability to recognize and reason about cognitive distortions, laying the groundwork for further exploration and refinement of our proposed framework.

I.2.1 Emotion Analysis for Cognitive Distortion and its Reasoning Task

Given the established relationship between cognitive distortion and emotion (Singh et al., 2023), we delve into the interplay between reasoning and emotion. The results depicted in Table 3 validate our initial hypothesis, demonstrating a discernible correlation between reasoning and emotional states. However, owing to space limitations, we were unable to include the detailed results of emotion analysis in the main paper.

To address this omission, we present the comprehensive findings regarding emotions in Table 6. This table offers a detailed breakdown of the emotional states associated with various types of reasoning. Each entry in the table provides insights into the emotional nuances underlying different forms of cognitive distortion reasoning, shedding light on the complex interrelationship between cognition and emotion.

By examining the emotional aspect alongside reasoning, we gain a deeper understanding of the cognitive processes involved in generating responses related to cognitive distortions. This holistic approach enables us to elucidate the intricate dynamics between cognitive distortion and emotional states, contributing to a more comprehensive analysis of the phenomenon.

In summary, the inclusion of emotion analysis complements our investigation into reasoning, enriching our understanding of the cognitive and affective dimensions of cognitive distortions. These findings collectively contribute to advancing our knowledge of the interplay between cognition and emotion in the context of mental health.

.

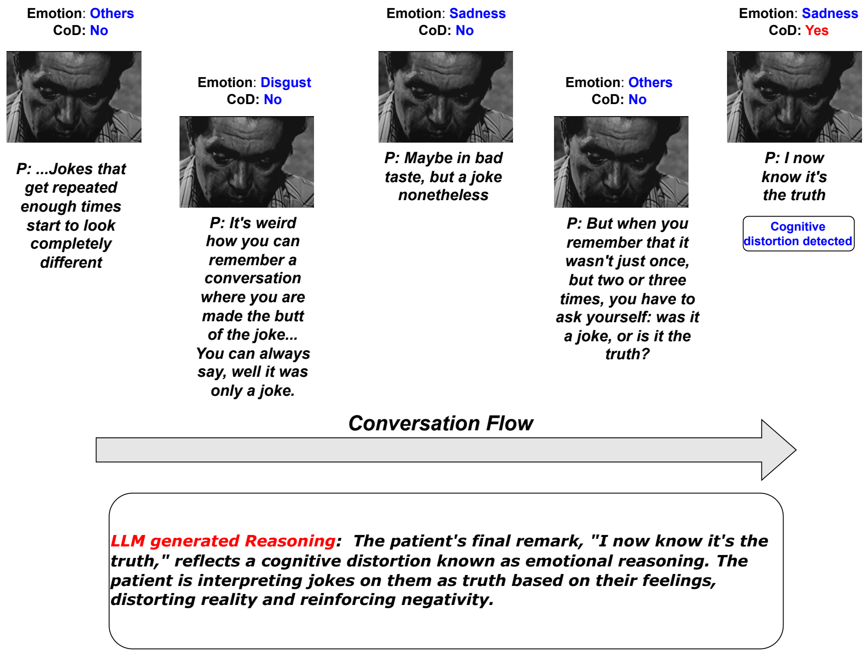


Figure 10: An example of reasoning generated by ZS-CoDR, with Alpaca LLM. The model is able to even generate what type of cognitive distortion is present, as in this "emotional reasoning" and explains it with respect to the patient's utterance.

Baseline	**F1^{ED}%**	**Acc^{ED}%**
DialogueRCN(Hu et al., 2021)	64.64	65.77
Bi-Direction RNN(Raheja and Tetreault, 2019)	63.49	67.03
One-LLM(Han et al., 2023)	72.64	76.53
SK+ZS Classifier(Zhang et al., 2019)	68.63	69.97
ZeroNLG(Yang et al., 2024)	70.53	74.72
ZS-CoDR(Proposed$_{LLaMA-7B}$ $+ EMOCA$)	72.77	**79.91**
ZS-CoDR(Proposed$_{LLaMA-7B}$)	**73.33**	**78.26**

Table 6: Automatic Evaluation Results for Emotion Detection. Where ED: Emotion Detection

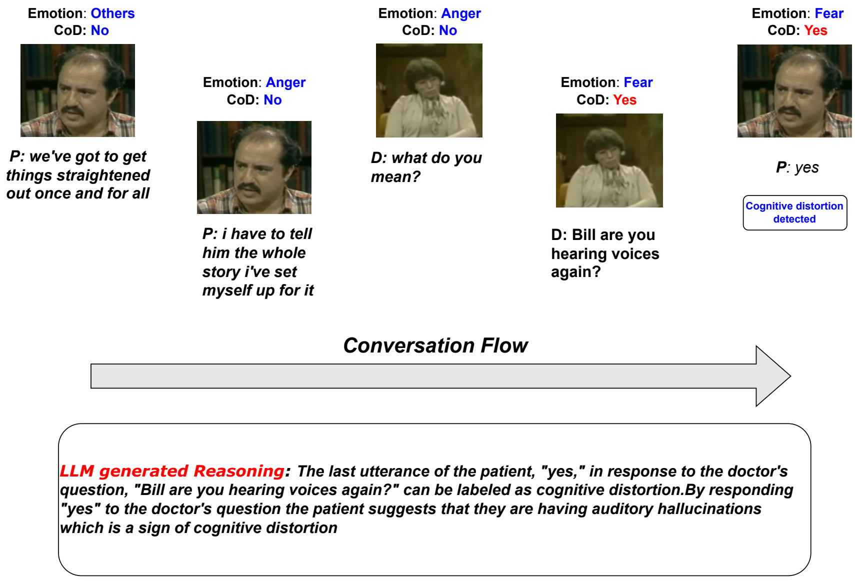


Figure 11: An example of reasoning generated by ZS-CoDR, with Vicuna LLM. The reasoning generated shows that the model could understand the context clearly, as it can deduce that the patient's response to the doctor's question about hearing voices signifies auditory hallucinations.

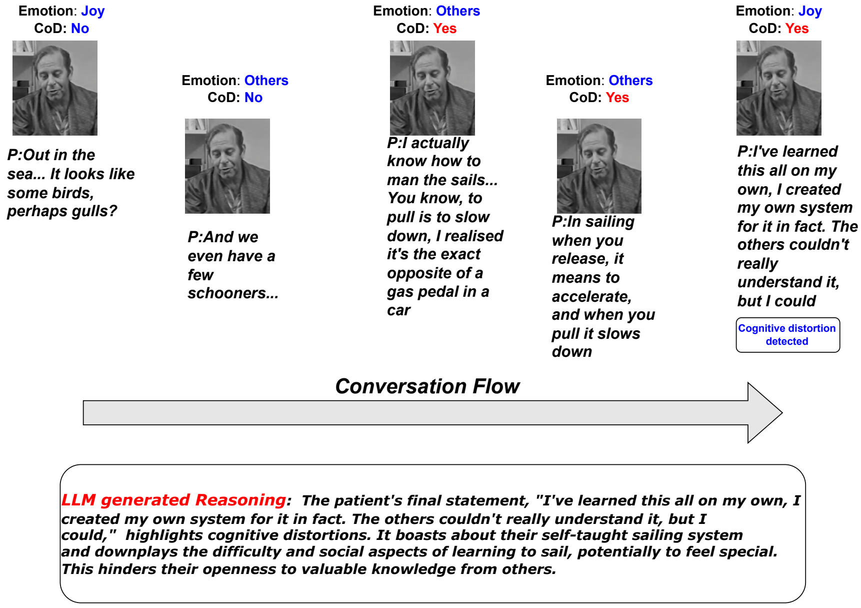


Figure 12: An example of reasoning generated by ZS-CoDR, with DOLLY LLM. Although the reasoning sets up the basis for cognitive distortion and utilize the emotion information, it does not explicitly mention cues in the context such as"others couldn't..." which reflect the patient's thought process.

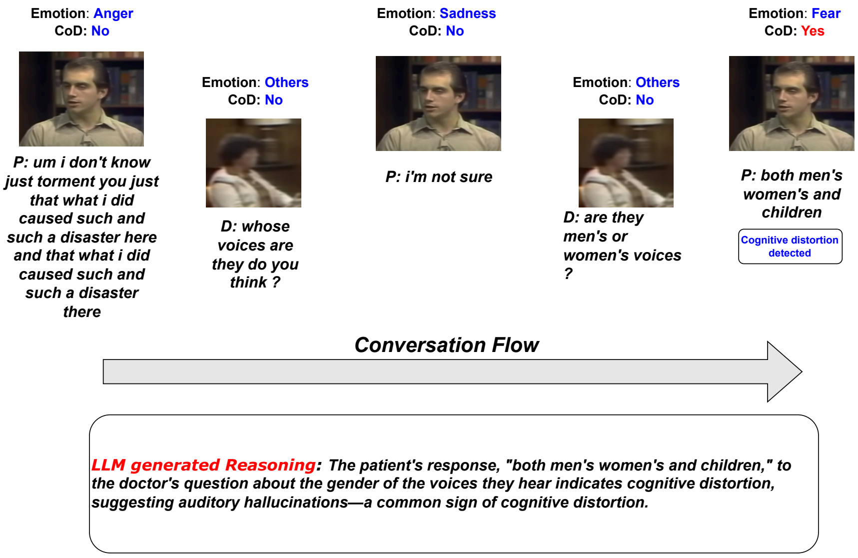


Figure 13: An example of reasoning generated by ZS-CoDR, with BLOOM LLM. Although the reasoning mentions the presence of auditory hallucinations, the reasoning as a whole is not very detailed compared to other LLMs.

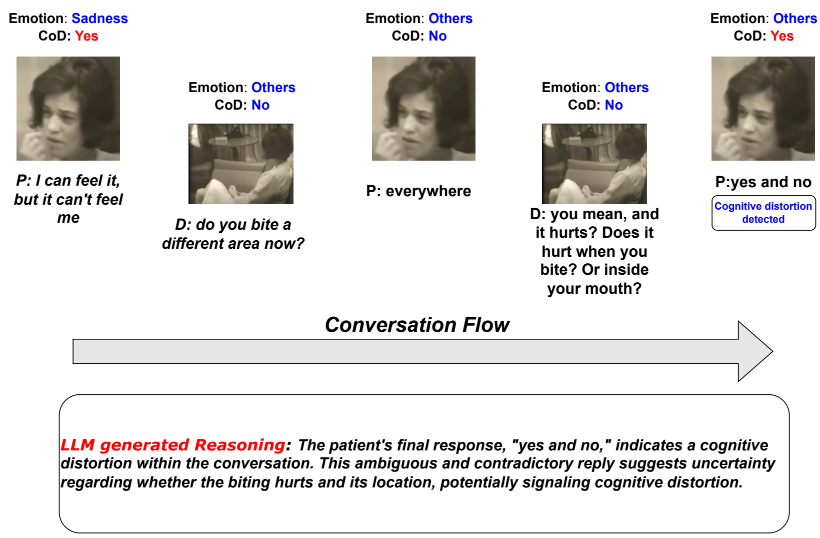


Figure 14: An example of reasoning generated by ZS-CoDR, with StableLM LLM. The reasoning contains how the patient's ambiguous response to the doctor's question reflects cognitive distortion.

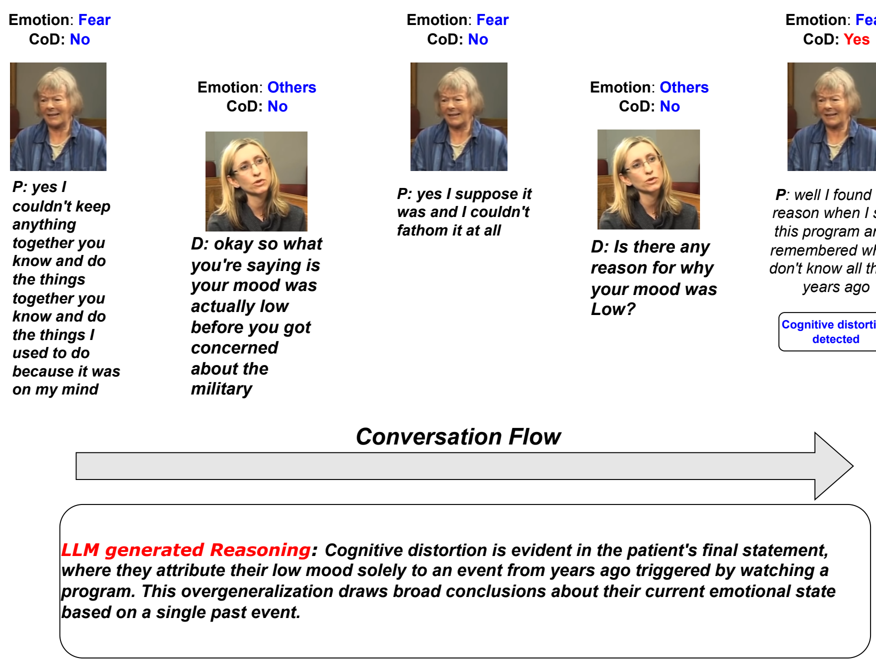


Figure 15: An example of reasoning generated by ZS-CoDR, with XLNet LLM. The reasoning is less coherent with the context compared to other LLMs. However, its able to utilize the emotion label to generate the reasoning, by mentioning that the "patient is triggered".

LLM	Conversation with CoD label	Reasoning
OPT	P: I don't have the will to do anything D: What would you like to do? P: all day, I just think P: I ruminate P: I deserve to be punished	The patient's feeling of deserving punishment is likely a distorted thought pattern. They blame themselves too much (taking full responsibility) and see things as extremes (punishment or nothing) instead of considering a range of possibilities.
LLaMA	P: I don't have the will to do anything D: What would you like to do? P: all day, I just think P: I ruminate P: I deserve to be punished	"The patient's belief that they deserve punishment appears to be a distorted pattern of thinking. They tend to overly blame themselves, assuming complete responsibility, and often view situations in black-and-white terms, without considering the possibility of other outcomes or perspectives."
BLOOM	P: I don't have the will to do anything D: What would you like to do? P: all day, I just think P: I ruminate P: I deserve to be punished	The patient's sense of deserving punishment may stem from a distorted cognitive pattern. They tend to excessively self-blame, assuming complete responsibility, and perceive situations dichotomously, either punishment or nothing, without exploring a spectrum of potential outcomes
MPT	P: I don't have the will to do anything D: What would you like to do? P: all day, I just think P: I ruminate P: I deserve to be punished	The patient's belief that they merit punishment likely indicates a twisted pattern of thinking. They excessively fault themselves and perceive situations in stark contrasts , neglecting to explore various options.
Alpaca	P: I don't have the will to do anything D: What would you like to do? P: all day, I just think P: I ruminate P: I deserve to be punished	The patient's conviction that they're worthy of punishment suggests a twisted way of thinking. They're excessively pointing the finger at themselves and painting situations in black and white, ignoring the rainbow of possibilities.
Vicuna	P: I don't have the will to do anything D: What would you like to do? P: all day, I just think P: I ruminate P: I deserve to be punished	The patient's belief that they deserve punishment could indicate a twisted way of thinking. They're overly harsh on themselves, shouldering all the blame.
Dolly	P: I don't have the will to do anything D: What would you like to do? P: all day, I just think P: I ruminate P: I deserve to be punished	The patient's belief that they deserve punishment indicates a distorted cognitive process. They excessively self-blame, assuming complete responsibility, and tend to view situations in binary terms, neglecting the nuances of potential outcomes
StableLM	P: I don't have the will to do anything D: What would you like to do? P: all day, I just think P: I ruminate P: I deserve to be punished	The patient's perception of deserving punishment suggests a cognitive distortion. They excessively attribute blame to themselves, assuming full responsibility, and tend to view situations in binary terms, rather than exploring a spectrum of possibilities.
XLNet	P: I don't have the will to do anything D: What would you like to do? P: all day, I just think P: I ruminate P: I deserve to be punished	The patient's unwavering sense of deserving punishment hints at a twisted cognitive labyrinth. They heap blame upon themselves relentlessly.
T5	P: I don't have the will to do anything D: What would you like to do? P: all day, I just think P: I ruminate P: I deserve to be punished	The patient's steadfast conviction of deserving punishment hints at a labyrinth of twisted cognition. They voraciously devour self-blame, indulging in it like a gourmet meal, unable to escape the suffocating tendrils of their own guilt.

Table 7: Comparison of reasoning generated by different LLMs for the same conversations.

LLM	Conversation with CoD label	Reasoning
GPT3.5	P: I don't have the will to do anything D: What would you like to do? P: all day, I just think P: I ruminate P: I deserve to be punished	"I deserve to be punished," the patient's final statement, reflects cognitive distortion as the patient is acknowledging personal accountability for their circumstances and ascribing it to merited retribution. This is indicative of assuming responsibility for unfavorable outcomes which in excessive could lead to guilt and self-reproach. Additionally, the patient assumes that they should be punished, which is reflects extreme thinking, without taking into account other viewpoints, which is typical sign of distorted thinking.
GPT3	P: I don't have the will to do anything D: What would you like to do? P: all day, I just think P: I ruminate P: I deserve to be punished	"I deserve to be punished," shows cognitive distortion because they're taking personal blame for their situation and believing they deserve punishment for it. This suggests they're accepting responsibility for negative outcomes, which could lead to feelings of guilt and self-blame if unchecked. Furthermore, they're not considering other perspectives, which is a common feature of distorted thinking.
GPT2	P: I don't have the will to do anything D: What would you like to do? P: all day, I just think P: I ruminate P: I deserve to be punished	"I deserve to be punished," demonstrates cognitive distortion as they are holding themselves personally accountable for their situation and feeling deserving of punishment. This indicates a tendency to accept blame for negative outcomes, potentially leading to feelings of guilt and self-blame if not addressed.

Table 8: Comparison of reasoning generated by different LLMs for the same conversations.

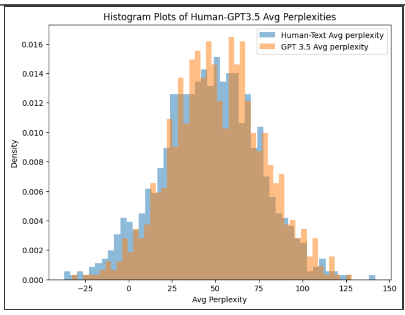
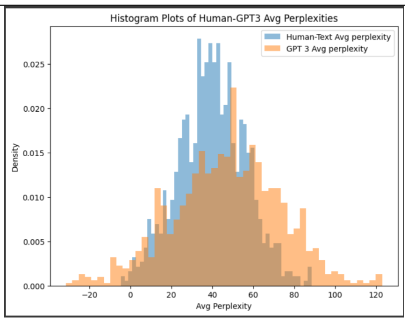
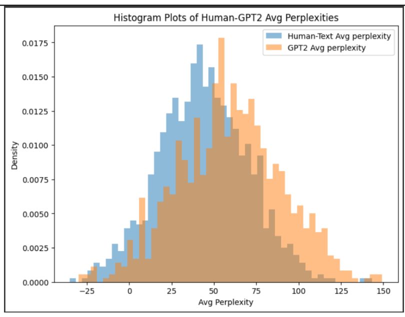
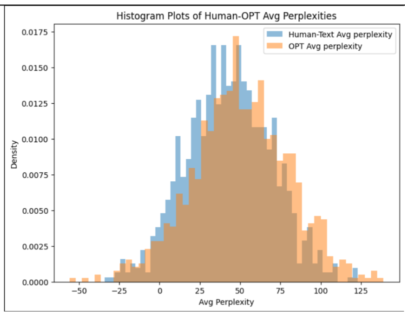
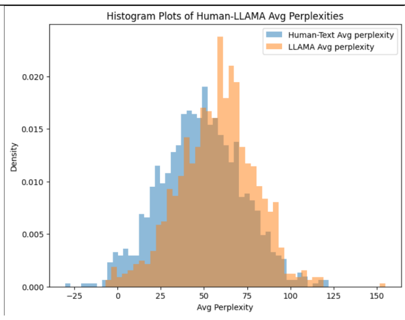
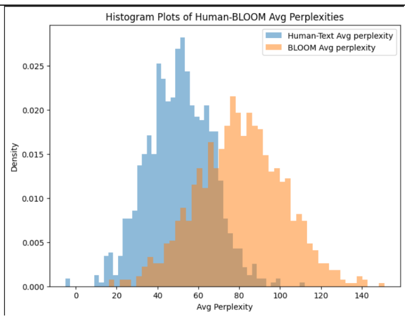
LLM	Perplexity Histogram
GPT 3.5	
GPT3	
GPT2	
OPT	
LLAMA	
BLOOM	

Table 9: Avergage Perplexity histogram plots comparison between human-annotated text and different LLMs

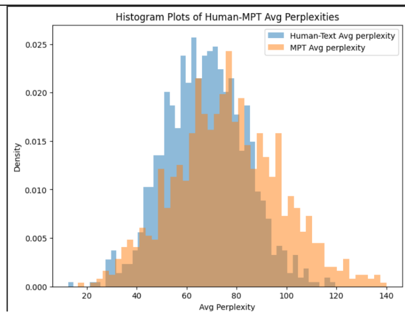
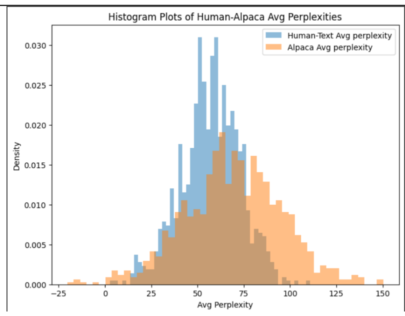
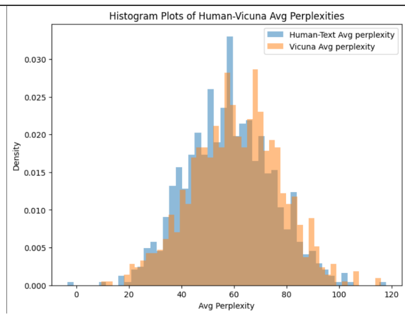
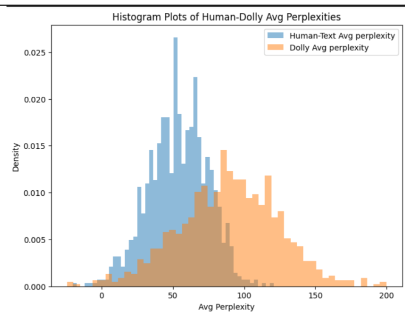
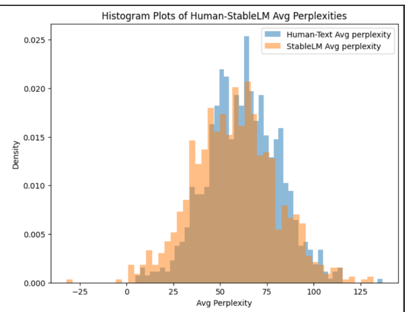
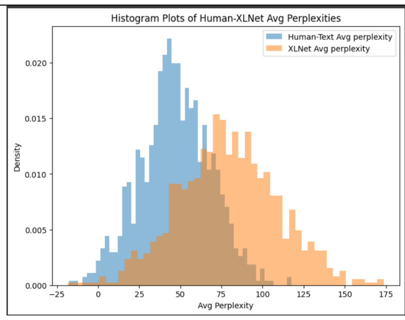
LLM	Perplexity Histogram
MPT	
Alpaca	
Vicuna	
Dolly	
StableLM	
XLNet	

Table 10: Avergage Perplexity histogram plots comparison between human-annotated text and different LLMs