From Span Extraction to Classification: A Multi-step Framework for Cognitive Distortion Analysis

Manh-Cuong Phan[1], **Thi-Ngoc-Phuong Nguyen**[1], **Huu-Loi Le**[2], **Huy-The Vu**[1], **Hajime Hotta**[3], and **Minh-Tien Nguyen**[1]*
[1] Hung Yen University of Technology and Education, Hung Yen, Vietnam.
cuongpm@spkt.edu.vn; {nguyenngocphuong, thevh, tiennm}@utehy.edu.vn
[2] AI Academy Vietnam, Vietnam.
loilh@aiacademy.edu.vn
[3] Hajime Institute, Kuala Lumpur, Malaysia.
hotta@hajime.institute

Abstract

Cognitive distortions (CDs) are biased thought patterns linked to conditions like depression and anxiety. Identifying CDs in therapy conversations is crucial for mental health support. Different from prior work that used rule-based and supervised methods but struggled with contextual understanding and data sparsity, we introduce a multi-step deep learning framework that first detects distortions, then leverages a machine reading comprehension module to extract distortion spans, and finally classifies their types. The task uses a machine reading comprehension module to extract distortion spans for noisy reduction, followed by a classifier to identify their types. Experimental results on a publicly available benchmark dataset, which has been widely adopted in prior studies, show that our framework achieves superior performance compared to strong baselines in distortion detection, span extraction, and classification.

1 Introduction

Cognitive distortions (CDs) refer to biased or irrational thought patterns that are often associated with various psychological disorders, such as depression, anxiety, and post-traumatic stress disorder (Beck, 2020). These distortions manifest in everyday conversations, e.g., patients and therapists, and can significantly hinder the therapeutic process (Beutel et al., 2019). Detecting and classifying cognitive distortions accurately within such interactions are crucial for multiple purposes, including practical applications such as enhancing therapeutic interventions (Chen et al., 2023b) and personalized mental health care (Shreevastava and Foltz, 2021), as well as research directions such as probing the reasoning ability of Large Language Models (LLMs) (Chen et al., 2023b; Wang et al., 2024; Lim et al., 2024).

*Corresponding Author.

Traditional methods for detecting cognitive distortions rely heavily on manual annotation and rule-based systems, which are limited in their scalability and ability to capture the subtle nuances of natural language (Shreevastava and Foltz, 2021). For example, Shreevastava and Foltz 2021 introduced a supervised learning framework for detecting CDs in patient-therapist interactions, leveraging feature engineering. Despite its success, this approach faces a major challenge relating to the use of the entire input (full speech text) that may add noisy information to CD models.

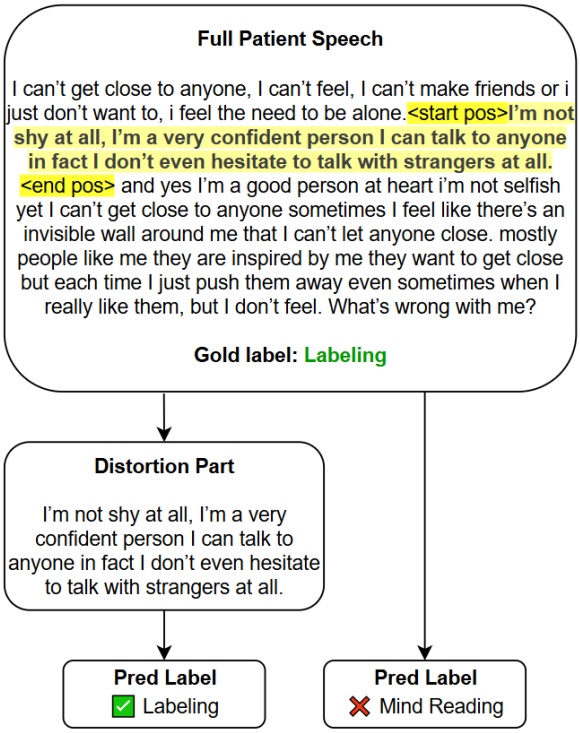

Figure 1: Example of cognitive distortion span extraction and classification. The full patient speech includes a distortion span (<start pos> and <end pos>). While our model correctly predicts the distortion type based on the span, Shreevastava and Foltz (2021) using the full speech incorrectly classifies it as "Mind Reading".

Let us take Figure 1 as an example. The patient's

speech contains a distortion span marked by start and end positions. When the model makes a prediction based on the full patient's speech, it incorrectly classifies the distortion as *Mind Reading* (Shreevastava and Foltz, 2021), likely due to the presence of unrelated or misleading context. However, when focusing solely on the distortion span, our model correctly identifies the distortion type as *Labeling*. This demonstrates the importance of isolating the distorted segment for accurate classification. Based on this observation, we argue that the performance of CD models can be improved by using distortion information extracted from the patient's speech.

This paper introduces a multi-step framework that takes advantage of span extraction for distortion detection. The framework consists of three main steps: (1) cognitive distortion detection, (2) distortion span extraction, and (3) distortion classification. The core hypothesis is that essential signals for distortion detection are often localized within a small portion of the input context. To address this, the framework first pinpoints these critical spans via a span extraction module, followed by a classification module that categorizes the type of distortions. The span extraction module uses pre-trained language models (PLMs) (e.g., BERT) to capture fine-grained semantic dependencies and contextual cues, enhancing detection performance. The framework is evaludated on the benchmark dataset introduced by Shreevastava and Foltz (2021), and experimental results demonstrate significant improvements over previous methods across multiple metrics. This paper makes two main contributions as follows.

- It introduces a framework for detecting and classifying cognitive distortions from patient-therapist interactions. The framework includes three steps: distortion detection, distortion span extraction, and final distortion classification in a single pipeline.
- It conducts a comprehensive evaluation on a publicly available dataset, demonstrating clear improvements across multiple metrics such as accuracy, precision, recall, and F1-score.

2 Related Work

Cognitive distortions are closely associated with mental health disorders such as depression and anxiety (Beck, 2020). The automatic detection and classification of such distortions has garnered increasing attention, particularly with advancements in natural language processing (NLP) and LLMs.

Shreevastava and Foltz (2021) introduced one of the first publicly available datasets focusing on cognitive distortions in patient speech. They annotated 2,531 utterances with both binary (distorted or not) and multi-class (distortion type) labels. Their semi-supervised approach leveraged labeled and unlabeled data but faced challenges in generalizability and interpretability due to limited context and data scale. To enhance interpretability, Chen et al. (2023c) proposed the Diagnosis of Thought (DoT) prompting framework, which utilizes LLMs such as ChatGPT and GPT-4. The DoT approach follows a structured reasoning process that consists of three steps: subjectivity assessment, contrastive reasoning, and schema analysis. Their results show that this method outperforms zero-shot baselines and provides clinically meaningful explanations, as confirmed by evaluations from licensed therapists.

Building on this foundation, Lim et al. (2024) introduced the Extraction, Reasoning, and Debate (ERD) framework, which employs multi-agent reasoning among LLMs to reduce diagnostic bias and improve performance in multi-class classification. Singh et al. (2024) investigated multimodal LLMs that integrate textual, auditory, and visual signals to improve zero-shot detection of distortions in patient–doctor interactions. Lin et al. (2024) developed a Mandarin-language dataset containing parallel reframing examples, enabling models to both detect distortions and suggest positive alternatives grounded in psychological theory.

In terms of scalability, Kim and Kim (2025) introduced KoACD, a large-scale dataset in Korean focusing on adolescent populations, and Babacan et al. (2025) leveraged GPT-4 to generate synthetic training data for cognitive distortion classification. While these datasets contribute to broader coverage across languages and domains, they are less applicable to our setting, which targets English patient–therapist interactions. Therefore, in this paper we evaluate our framework on the benchmark dataset introduced by Shreevastava and Foltz (2021).

While sharing the goal of cognitive distortion detection with Shreevastava and Foltz (2021) and Chen et al. (2023c), our work differs by adding a span extraction stage formulated as a machine reading comprehension (MRC) task. This step isolates distortion-relevant text, reducing noise and improving classification accuracy.

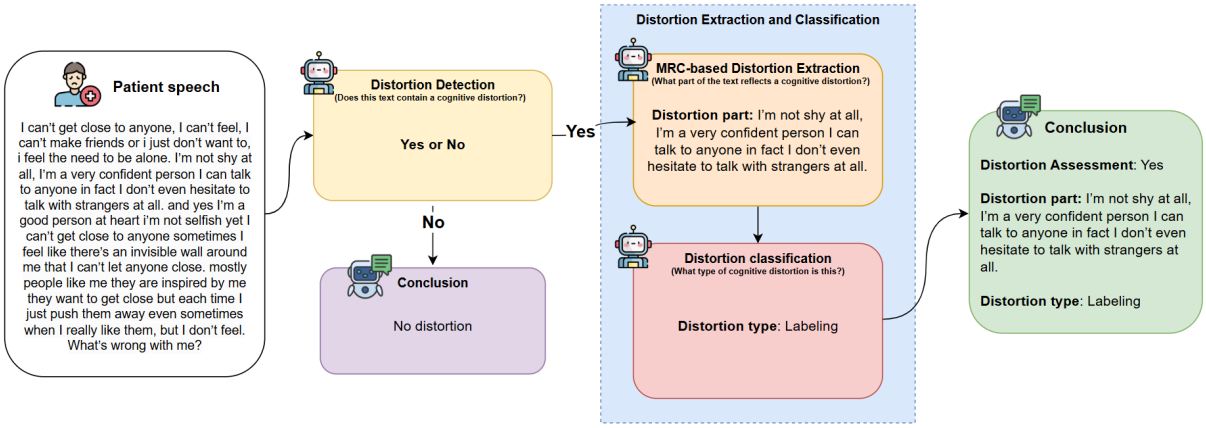

Figure 2: The proposed framework using machine learning techniques.

3 Methodology

3.1 Problem Statement

Cognitive distortions are irrational thought patterns that negatively impact mental health. Given an input text x (n tokens) (a conversation between a patient and a therapist), our goal is to (1) assess whether x contains any CD, (2) extract the span(s) responsible for the distortion, and (3) classify each extracted span into one of K predefined categories. We formalize the overall problem of cognitive distortion analysis in patient–therapist conversations as three learning subtasks: cognitive distortion detection (Section 3.3), distortion span extraction (Section 3.4), and distortion type classification (Section 3.5). Each task is trained independently with its objective function.

3.2 Overview of the Framework

Figure 2 shows the proposed framework for distortion detection and classification. The proposed framework consists of three primary components: distortion detection, the extraction of cognitive distortions using span extraction, and the classification of these distortions into specific types. The system processes patient–therapist interaction texts through a multi-stage pipeline designed to identify and classify cognitive distortions.

3.3 Cognitive Distortion Detection

Detection is performed first to filter out distortion-free utterances. This reduces noise and search space, making span extraction and type classification more accurate and efficient.

The first step detects whether each patient utterance is distortion-free or not. To do that, we formulate the detection as a binary classification problem that involves learning a classifier $f_d : \mathcal{X} \rightarrow \{0, 1\}$ which determines whether an utterance x_i contains a cognitive distortion. The prediction is denoted as:

$$y_i = f_d(x_i; \theta)$$

The detection is done in three steps: pre-processing, feature representation, and classification.

Pre-processing The dataset used for training the model is first preprocessed. This involves cleaning the text, tokenizing it into smaller units (tokens), and converting it into a format suitable for input into the MRC model. Text normalization techniques, such as removing stop words and punctuation, are applied to improve the quality of the input data. We employ the `nltk` and `re` libraries for basic text preprocessing, such as lowercasing, punctuation removal, and stop-word filtering. For PLM-based models (e.g., BERT, RoBERTa, DeBERTa), we apply only minimal preprocessing. Specifically, we rely on the model-specific subword tokenization and encoding provided by the Hugging Face `AutoTokenizer`[1], with lowercasing applied only when using uncased variants. We avoid aggressive transformations such as stop-word removal or punctuation stripping, since these may discard semantically informative tokens and harm PLM performance.

Class conversion The original dataset contains multiple classes representing different types of cognitive distortions, such as *Labeling*, *Mind Reading*, *Catastrophizing*, etc. To make the dataset suitable for the task of **distortion detection**, we convert these original classes into two binary categories:

[1] https://huggingface.co/docs/transformers

Distortion and **No-distortion**. All samples that contain any specific type of cognitive distortion are grouped under the *Distortion* class, while the rest (including normal or undefined responses) are labeled as *No-distortion*. This conversion enables the use of binary classification models while preserving the key distinction between distorted and non-distorted content. Tables 1 and 2 shows the statistics of labels of original and converted classes.

Table 1: Original class distribution.

Original Class	Sample Count
No-distortion	933
Mind Reading	239
Overgeneralization	239
Magnification	195
Labeling	165
Personalization	153
Fortune-telling	143
Emotional Reasoning	134
Mental filter	122
Should statements	107
All-or-nothing thinking	100

Table 2: Binary class distribution.

Binary Class	Sample Count
Distortion	1597
No-distortion	933

Feature representation. In this work, we adopt three different methods to represent an input conversation x as contextual vectors. The first method is Bag-of-Word (BoW) that creates a dictionary on the whole corpus and then maps each input x_i to a fixed-size vector (Joachims, 1998). The second method is TF-IDF that focuses more on the importance of words in the corpus (Ramos, 2003).

The third feature representation method leverages PLMs such as BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), or DeBERTa (He et al., 2023).

This is because these PLMs were trained with a huge amount of data to capture contextual information of input tokens. Given an input x_i, the framework adds a new [CLS] token to x_i to form a new sequence: [CLS] x_i. This sequence is fed into PLMs, and we take the hidden representation of the [CLS] token from the final encoder layer to form $\mathbf{H}$. The vector $\mathbf{H}$ is then passed to a classification layer for prediction.

Classification Given contextual vectors from feature representation, classification uses two types of methods: traditional and PLMs. Traditional methods use BoW and TF-IDF features, and PLMs use the hidden vectors $\mathbf{H}$ for classification. Results are shown in Tables 4 and 5.

3.4 Distortion Span Extraction

Span extraction aims to identify one or more spans within the input x_i that correspond to distorted expressions. Each span is defined as $s_i = (\text{start}_i, \text{end}_i) \subset x_i$, and the set of all extracted spans (m spans) is represented as follows.

$$P = f_{se}(x; \theta), \quad \text{where } P = \{p_1, p_2, \ldots, p_m\}.$$

We observed that distortion cues are localized to small segments rather than the entire utterance (Shreevastava and Foltz, 2021). Different from prior work that fed the whole utterance x_i for the final classification, we argue that using the full utterance may introduce noise for classifiers. To mitigate this, the framework extracts distortion spans from each utterance x_i for the final classification.

The extraction is formulated as a MRC (Machine Reading Comprehension) problem due to two reasons. First, MRC models can be utilized to pinpoint specific segments within a patient's narrative that indicate distorted thinking. By framing the detection task as a question-answering problem, the model can be prompted with questions such as, "What part of the text reflects a cognitive distortion?" This formulation allows the framework to focus on extracting evidence-based segments that signify distorted thoughts (Nguyen et al., 2023; Chen et al., 2023a). Second, recent studies have shown the efficacy of MRC frameworks in clinical concept extraction, highlighting their potential in identifying nuanced psychological patterns, including cognitive distortions (Chen et al., 2023a).

Given an input utterance $x_i = \{w_1, w_2, ..., w_n\}$ consisting of n tokens, we obtain its contextualized representations $\mathbf{H} = \{\mathbf{h}_1, \mathbf{h}_2, ..., \mathbf{h}_n\}$ from the PLM encoder. The span extraction module then predicts a set of distortion spans $P = \{p_1, p_2, ..., p_m\}$, where each span s_i is represented by its start and end positions ($S, E \in \mathbb{R}^c$) following the standard BERT QA formulation (Devlin et al., 2019).

The prediction uses a softmax over all hidden features $\mathbf{H}$ to predict the start or end positions of a

token j^{th} for the answer span p_i as follows.

$$p_{si}^{j} = \frac{\exp(h_i^{j\top} S)}{\sum_{i'} \exp(h_{i'}^{j\top} S)}; \quad p_{ei}^{j} = \frac{\exp(h_i^{j\top} E)}{\sum_{i'} \exp(h_{i'}^{j\top} E)}$$

The start and end positions of answer span p_i were calculated as follows.

$$s^i = argmax_i(p_{si}^j); \qquad e^i = argmax_i(p_{ei}^j)$$

For training, we utilize annotated datasets in which each distortion instance is labeled with its corresponding text span. These gold spans serve as supervision signals for the model to learn start and end token positions using cross-entropy loss. Table 6 shows results of the extraction.

3.5 Cognitive Distortion Classification

The final step is to predict distortion types of each input utterance x_i using the extracted spans in Section 3.4. For each patient utterance x_i, the span s_i identified in the previous step is classified by a model f_{cl} into one of K distortion categories:

$$c_i = f_{cl}(s_i; \theta), \quad c_i \in \{1, 2, \ldots, K\}.$$

Followed by Section 3.3, the framework uses three types of feature representation: BoW, TF-IDF, and contextual vectors from PLMs (BERT, RoBERTa and DeBERTa). The final classification also follows methods in Section 3.3.

To improve classification accuracy, we fine-tune pre-trained language models (BERT, RoBERTa, DeBERTa) on labeled spans. Each span s_i is passed through a classification head (a linear layer and softmax) to predict distortion types. The models are optimized using cross-entropy loss and AdamW optimizer (Loshchilov and Hutter, 2019), with early stopping applied based on validation performance.

4 Experimental Settings

4.1 Dataset

Experiments were conducted on the annotated dataset introduced by Shreevastava and Foltz (2021). It contains 2,531 utterances extracted from real-world patient-therapist sessions. Each utterance is annotated with binary labels (distortion or non-distortion). If distorted, it is further classified into one of ten predefined cognitive distortion types (Table 3). The dataset also includes start and end positions of distortion spans. It was split into training and test sets using an 80/20 ratio, ensuring consistent distribution across distortion categories.

4.2 Settings

Our proposed framework consists of three core components, each trained under the following settings.

Distortion detection. We treat distortion assessment as a binary classification task. Following Shreevastava and Foltz (2021), we experiment with five machine learning models: **Logistic Regression** (Hosmer et al., 2013), **Support Vector Machines (SVM)** (Cortes and Vapnik, 1995), **Decision Tree** (Safavian and Landgrebe, 1991), **k-NN** (Cover and Hart, 1967), and **MLP** (Rumelhart et al., 1986). Each classifier is trained with two types of feature representations: Bag-of-Words (BoW) and TF-IDF. In addition, we fine-tune **BERT**, **RoBERTa**, and DeBERTa (Devlin et al., 2019; Liu et al., 2019; He et al., 2023) for binary classification using Hugging Face's `Transformers` library (Wolf et al., 2020). For fine-tuned PLMs, we use a batch size of 16, learning rate of 2e-5, AdamW optimizer, and train for 5 epochs.

Distortion span extraction. This task is formulated as a span-based question answering problem, following the MRC paradigm. Given an utterance and a query such as "Which part of the text reflects a cognitive distortion?", the model outputs the relevant span. We fine-tune **BERT**, **RoBERTa**, and **DeBERTa** using the Hugging Face `QA pipeline`[2]. Training is performed with a batch size of 12, learning rate of 3e-5, AdamW optimizer, and 3 epochs.

Cognitive distortion type classification. After extracting distortion spans, the framework classifies them into specific distortion categories. We employ the same feature representations and classifiers as distortion detection. For fine-tuned PLMs, we use a batch size of 16, learning rate of 2e-5, and 5 epochs.

4.3 Evaluation Metrics

Each subtask is evaluated using task-appropriate metrics as follows. **Distortion detection** uses F1 score for the positive class (i.e., distorted utterances) (Shreevastava and Foltz, 2021). **Span extraction** uses Exact Match (EM) and token-level F1 score, following common practice in span-based QA tasks (Rajpurkar et al., 2016). **Cognitive distortion classification** uses accuracy and weighted

[2] https://huggingface.co/docs/transformers/en/main_classes/pipelines#transformers.QuestionAnsweringPipeline

Table 3: Common cognitive distortion types and example speech (Beck, 2020; Shreevastava and Foltz, 2021).

Cognitive Distortion Type	Interpretation	Example Distorted Speech
Personalization	Personalizing or taking up the blame for a situation that involved many factors outside the person's control.	My son is pretty quiet today. I wonder what I did to upset him.
Mind Reading	Suspecting what others are thinking or the motivations behind their actions.	My house was dirty when my friends came over; they must think I'm a slob!
Overgeneralization	Drawing major conclusions based on limited information.	Last time I was in the pool I almost drowned; I am a terrible swimmer and should not go into the water again.
All-or-nothing thinking	Seeing a situation as either black or white with no middle ground.	If I cannot get my Ph.D., then I am a total failure.
Emotional reasoning	Letting feelings override factual evidence.	Even though Steve is here at work late every day, I know I work harder than anyone else at my job.
Labeling	Assigning a fixed label to oneself or others without deeper examination.	My daughter would never do anything I disapproved of.
Magnification	Emphasizing the negative or minimizing the positive aspects of a situation.	My professor said he made some corrections on my paper, so I know I'll probably fail the class.
Mental filter	Focusing only on the negatives of a situation.	My husband says he wishes I was better at housekeeping, so I must be a lousy wife.
Should statements	Creating rigid rules about how one or others should behave.	I should get all A's to be a good student.
Fortune-telling	Predicting that things will turn out badly without evidence.	I was afraid of job interviews so I decided to start my own thing.

F1 score to account for class imbalance across ten categories.

5 Results and Discussion

5.1 Performance Comparison

This section reports the comparison of the proposed framework to baselines for three problems: distortion detection, span extraction, and distortion classification. All reported metrics are computed on the held-out test set.

5.1.1 Distortion detection

Performance with traditional methods To evaluate the effectiveness of different feature representations for cognitive distortion detection, we experimented with various linguistic and semantic feature sets. These include Sentence Embeddings using SIF (Arora et al., 2017), BERT-based embeddings (Reimers and Gurevych, 2019), psycholinguistic features from LIWC (Pennebaker et al., 2001), Part-of-Speech (POS) tags (Toutanova et al., 2003), and combinations thereof. The results of these representations are derived from original papers (Shreevastava and Foltz, 2021) for fair comparison. Additionally, we compared these against our features: BoW and TF-IDF (Section 3.3).

As shown in Table 4, simple lexical representations such as BoW and TF-IDF consistently outperform more complex embedding-based features in this task. TF-IDF achieves the strongest overall results across most classifiers, while BoW also performs competitively, particularly with certain tree-based methods. Among the learned embedding approaches, combinations incorporating psycholinguistic features (e.g., LIWC) yield moderate improvements, highlighting that such features still hold value for cognitive distortion assessment.

Table 4: Evaluation of traditional methods. LR: Logistic Regression, DT: Decision Tree. Reported metric is F1. † Results copied from Shreevastava and Foltz (2021).

Feature	LR	SVM	DT	k-NN	MLP
SIF †	0.75	0.77	0.65	0.74	0.73
BERT †	0.74	0.79	0.67	0.75	0.70
LIWC †	0.77	0.78	0.67	0.76	0.77
POS †	0.73	0.77	0.66	0.75	0.72
BERT+LIWC †	0.74	0.76	0.64	0.75	0.74
BoW (our)	0.77	0.80	**0.71**	0.71	0.74
TF-IDF (our)	**0.81**	**0.81**	0.70	**0.79**	**0.78**

These findings suggest that, for this domain, sparse and interpretable lexical features can be more effective than dense contextual embeddings, possibly due to the distinct and recurring linguistic markers associated with distorted thinking.

Comparison of PLMs Table 5 shows the detection results using BERT, RoBERTa and DeBERTa. Compared to traditional methods in Table 4, the detection using PLMs obtains better performance. This is because PLMs were trained with a huge amount of data. When fine-tuning for downstream tasks, they usually produce better accuracy than traditional methods, which have much smaller model sizes.

Table 5: Distortion detection with PLMs.

Models	Precision	Recall	F1
BERT-base	0.76	0.89	0.82
BERT-large	0.76	0.91	0.83
RoBERTa-base	0.76	**0.93**	**0.84**
RoBERTa-large	0.77	**0.93**	**0.84**
DeBERTa-v3-base	**0.78**	0.91	**0.84**
DeBERTa-v3-large	**0.78**	0.91	**0.84**

5.1.2 Span extraction

Table 6 shows that model capacity and architectural design both play an important role in distortion span extraction. Larger variants of BERT and RoBERTa consistently outperform their base counterparts, suggesting that increased parameterization enhances the ability to capture fine-grained cues. Moreover, DeBERTa-v3 achieves the strongest overall performance, indicating that architectural refinements such as disentangled attention further improve span identification beyond mere scaling.

Table 6: Evaluation of distortion extraction.

Models	EM	F1
BERT-base	72.92	84.11
BERT-large	77.27	87.46
RoBERTa-base	75.89	85.23
RoBERTa-large	77.08	85.77
DeBERTa-v3-base	78.26	88.05
DeBERTa-v3-large	**78.54**	**88.47**

These findings support the feasibility of formulating cognitive distortion extraction as a machine reading comprehension problem and demonstrate that contextualized language models are effective in capturing fine-grained psychological patterns.

5.1.3 Cognitive distortion classification

This section shows the performance of distortion classification in two settings: using two steps (span extraction and classification) and the full pipeline. This design isolates classifier performance from upstream errors and reveals how detection mistakes propagate to final classification.

Step-wise evaluation with traditional methods The classification task was performed on two settings: the full speech text and the extracted distorted spans (distorted parts). The full speech text uses the whole utterance for classification. The extracted distorted spans use extracted spans from Section 3.4 for classification. We also include gold labels to observe the gaps between extracted spans and gold-labeled spans.

Table 7 shows consistent trends across traditional and transformer-based models. Using entire utterances leads to lower performance due to noise and irrelevant content, whereas gold-standard spans yield clear improvements. Automatically extracted spans perform close to the gold setting, indicating that the extraction step effectively reduces noise while retaining task-relevant information.

Table 7: Results for cognitive distortion classification with step-wise and full pipeline. † Results copied from Shreevastava and Foltz (2021).

Speech Part	Methods	BoW		TF-IDF	
		Acc	F1	Acc	F1
Full speech	K-NN †	–	–	–	0.24
	Logistic Reg.	0.23	0.23	0.24	0.19
	SVM	0.19	0.19	0.28	0.24
	Decision Tree	0.13	0.13	0.16	0.16
	K-NN	0.16	0.16	0.19	0.20
	MLP	0.23	0.23	0.25	0.24
		Acc		F1	
	BERT-base	0.27		0.25	
	BERT-large	0.29		0.28	
	RoBERTa-base	0.28		0.28	
	RoBERTa-large	0.30		0.33	
	DeBERTa-base	0.25		0.25	
	DeBERTa-large	0.27		0.28	
Distortion span extraction (gold)	Logistic Reg.	0.32	0.32	0.35	0.31
	SVM	0.32	0.32	0.35	0.35
	Decision Tree	0.20	0.20	0.17	0.18
	K-NN	0.16	0.15	0.20	0.20
	MLP	0.32	0.32	0.32	0.32
		Acc		F1	
	BERT-base	0.41		0.42	
	BERT-large	0.47		0.47	
	RoBERTa-base	0.43		0.44	
	RoBERTa-large	0.47		0.48	
	DeBERTa-base	0.43		0.43	
	DeBERTa-large	0.45		0.47	
Distortion span extraction (ours)	Logistic Reg.	0.31	0.31	0.32	0.28
	SVM	0.31	0.31	0.31	0.30
	Decision Tree	0.18	0.19	0.19	0.19
	K-NN	0.16	0.15	0.23	0.23
	MLP	0.30	0.29	0.29	0.28
		Acc		F1	
	BERT-base	0.38		0.39	
	BERT-large	0.41		0.42	
	RoBERTa-base	0.43		0.44	
	RoBERTa-large	**0.45**		**0.46**	
	DeBERTa-base	0.42		0.42	
	DeBERTa-large	0.41		0.43	
Full pipeline	Logistic Reg.	0.28	0.29	0.29	0.28
	SVM	0.25	0.24	0.26	0.23
	Decision Tree	0.14	0.17	0.16	0.19
	K-NN	0.11	0.13	0.21	0.21
	MLP	0.25	0.28	0.26	0.28
		Acc		F1	
	BERT-base	0.33		0.36	
	BERT-large	0.36		0.40	
	RoBERTa-base	0.36		0.39	
	RoBERTa-large	0.38		0.42	
	DeBERTa-base	0.38		0.41	
	DeBERTa-large	0.40		0.42	

Evaluation with PLMs For transformer-based models, a similar trend is observed. Full-speech inputs result in the lowest performance, gold-standard spans lead to the largest gains, and automatically extracted spans maintain performance levels close to the gold spans. These results reinforce the hypothesis that focusing on distorted segments is an effective strategy for improving cognitive distortion classification.

Full pipeline evaluation We further evaluate the full pipeline performance, where span extraction and classification are executed in sequence. The lower part of Table 7 shows that transformer-based models still outperform traditional ones. Compared to classification on gold spans, these results of the full pipeline show competitive performance, which is still better than directly using full text. The scores of traditional classifiers using the full pipeline are better than those of using full speech. This confirms the contribution of distortion span extraction. However, the performance of methods using distortion span extraction is sill better than that of using full pipeline due to error accumulation.

5.1.4 Discussion with LLM-based models

Here we compare our framework, which relies on relatively small models, with LLMs that contain orders of magnitude more parameters.

Table 8: LLM-based models results. ($\star$) results copied from Chen et al. (2023c). Numbers in subscript denote the standard deviation over five runs.

Methods	Distortion Detection (F1)	Distortion Classification (Weighted F1)
Full training*	75.00	24.00
Vicuna*	$73.81_{0.95}$	$11.23_{0.78}$
ChatGPT*	$73.47_{0.58}$	$19.24_{1.00}$
ChatGPT + ZCoT*	$77.10_{1.21}$	$20.21_{1.02}$
ChatGPT + DoT*	$81.19_{0.11}$	$22.25_{0.70}$
GPT-4*	$83.04_{0.51}$	$33.86_{0.83}$
GPT-4 + ZCoT*	$81.97_{1.21}$	$33.22_{1.36}$
GPT-4 + DoT*	$82.77_{0.81}$	$34.64_{1.40}$
RoBERTa-base (our)	**84.00**	**44.00**

The higher scores of our method in Table 8 do not imply a direct superiority over the GPT-based approaches, but rather reflect the advantage of task-specific fine-tuning. This is because while the GPT-based methods in Table 8 operate in a zero-shot setting, our RoBERTa-base model is fine-tuned on the target dataset. This table is included to provide an additional observation on the performance gap between large, general-purpose LLMs without training on domain data and smaller, fine-tuned transformer models.

5.2 Error Analysis

Table 9 presents two representative cases. In Case 1, the full-speech model is misled by **mind-reading cues** ("people like me, they are inspired by me"), while our span-based model focuses on **self-labeling expressions** ("i'm ...a very confident person", "i'm not shy at all"), leading to the correct prediction of `Labeling`. In Case 2, the utterance contains **threat amplification and consequence escalation** ("every little thing ...terrified", "physically can't sleep", "so afraid"), which indicate `Magnification`. However, the full-speech model is influenced by **generalization cues**, predicting `Overgeneralization`, and our span-based model is biased by **first-person pronouns** and self-reference, predicting `Personalization`.

Table 9: Examples of two representative cases in error analysis.

Case 1: Full-speech misclassified, span prediction correct	**Case 2: Both full-speech and span prediction misclassified**
Patient speech (shortened): "...mostly people like me, they are inspired by me ...but each time I just push them away ..."	*Patient speech (shortened):* "...every little thing is making me terrified ...unless I take benedryl I physically can't sleep ...minor hallucinations ..."
Extracted span: "i'm not shy at all ...i'm a very confident person ...i don't even hesitate to talk with strangers ..."	*Extracted span:* "I see the light flickering ...someone's there ...every little thing is making me terrified ...I physically can't sleep ..."
Gold: `Labeling`	*Gold:* `Magnification`
Full-speech prediction: `Mind Reading`	*Full-speech prediction:* `Overgeneralization`
Extracted span prediction: `Labeling`	*Extracted span prediction:* `Personalization`

6 Conclusions

This paper presented a multi-step framework for cognitive distortion analysis in patient–therapist dialogues. Unlike prior studies that operate on full utterances, the proposed method explicitly models distortions at the span level by formulating span extraction as a machine reading comprehension task. The framework first detects whether a cognitive dis-

tortion is present, extracts the corresponding text spans, and then classifies them into fine-grained distortion categories. By isolating distortion-relevant spans, the approach reduces irrelevant context and enhances both interpretability and classification accuracy.

Comprehensive experiments on a benchmark dataset demonstrated consistent improvements over strong baselines, including traditional feature-based classifiers and pre-trained language models applied directly to full utterances. Span-level modeling was shown to yield notable gains in multi-class classification, with transformer-based models such as RoBERTa and DeBERTa achieving state-of-the-art performance. These findings indicate that sub-utterance representations are more effective than coarse-grained utterance-level inputs, particularly in domains requiring nuanced semantic distinctions.

The study also contributes to the growing body of computational methods for supporting cognitive-behavioral therapy (CBT). By aligning the modeling pipeline with the way distortions are annotated in clinical practice, the proposed framework provides results that are both more accurate and more interpretable, making it a promising step toward real-world mental health applications.

Future research directions include extending the framework to multi-span extraction and label-aware span selection, validating the approach on multilingual and cross-domain datasets such as KoACD and synthetic corpora, and incorporating multimodal signals (e.g., prosody, facial expressions, dialogue context) to enhance robustness in naturalistic settings.

Overall, the findings establish span-based cognitive distortion detection as a promising research direction, bridging the gap between automated NLP techniques and clinically meaningful psychological constructs. This framework lays the foundation for more accurate, interpretable, and clinically applicable systems in computational mental health.

Limitations

While the proposed framework performs well on the benchmark dataset, it relies on supervised span annotations, which may be unavailable in some domains. Evaluation is limited to English patient–therapist interactions, leaving its applicability to other languages, cultures, and settings untested. The current approach also focuses solely on textual data, excluding multimodal cues such as prosody or facial expressions.

Ethics Statement

This work uses publicly available, anonymized datasets for research purposes. The system is intended to support, not replace, mental health professionals, and should not be used as a stand-alone diagnostic tool. Potential biases in predictions must be monitored to avoid misclassification and possible harm in clinical contexts.

Acknowledgements

The authors thank the anonymous reviewers for their valuable feedback and suggestions. Support from colleagues and institutions during this research is gratefully acknowledged.

References

Sanjeev Arora, Yingyu Liang, and Tengyu Ma. 2017. A simple but tough-to-beat baseline for sentence embeddings. *International Conference on Learning Representations (ICLR)*.

Erdem Babacan, Melis Kaya, and Tolga Ozdemir. 2025. Synthetic cognitive distortion data generation with gpt-4 for safer model training. *Fırat University Medical Bulletin*, 30(1).

Judith S Beck. 2020. *Cognitive behavior therapy: Basics and beyond*. Guilford Publications.

M. E. Beutel et al. 2019. Cognitive distortions and their role in the development of mental disorders. *European Psychiatry*, 56:48–54.

Danqi Chen et al. 2023a. Clinical concept extraction with machine reading comprehension models. *Journal of Artificial Intelligence in Medicine*, 118:103329.

Zhiyu Chen, Yujie Lu, and William Wang. 2023b. Empowering psychotherapy with large language models: Cognitive distortion detection through diagnosis of thought prompting. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 4295–4304.

Zhiyu Chen, Yujie Lu, and William Wang. 2023c. Empowering psychotherapy with large language models: Cognitive distortion detection through diagnosis of thought prompting. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 4295–4304, Singapore. Association for Computational Linguistics.

Corinna Cortes and Vladimir Vapnik. 1995. Support-vector networks. *Machine learning*, 20(3):273–297.

Thomas Cover and Peter Hart. 1967. Nearest neighbor pattern classification. *IEEE transactions on information theory*, 13(1):21–27.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2023. Debertav3: Improving deberta using electra-style pretraining with gradient-disentangled embedding sharing. *Preprint*, arXiv:2111.09543.

David W Hosmer, Stanley Lemeshow, and Rodney X Sturdivant. 2013. *Applied logistic regression*. John Wiley & Sons.

Thorsten Joachims. 1998. Text categorization with support vector machines: Learning with many relevant features. In *European conference on machine learning*, pages 137–142. Springer.

Jiwon Kim and Seunghyun Kim. 2025. Koacd: A large-scale korean dataset for adolescent cognitive distortion detection. arXiv preprint arXiv:2505.00367.

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. *arXiv preprint arXiv:2403.14255*.

Xiaohui Lin, Jun Zhou, and Yuwei Wang. 2024. Positive reframing of distorted thoughts: A mandarin dataset and llm evaluation. In *Findings of ACL 2024*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1833–1844.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations (ICLR)*.

Minh-Tien Nguyen, Nguyen Hong Son, et al. 2023. Gain more with less: Extracting information from business documents with small data. *Expert Systems with Applications*, 215:119274.

James W. Pennebaker, Martha E. Francis, and Roger J. Booth. 2001. *Linguistic Inquiry and Word Count: LIWC*. Erlbaum Publishers.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. In *Proceedings of EMNLP*, pages 2383–2392.

Juan Ramos. 2003. Using tf-idf to determine word relevance in document queries. In *Proceedings of the first instructional conference on machine learning*, volume 242, pages 133–142.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *Preprint*, arXiv:1908.10084.

David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. 1986. Learning representations by back-propagating errors. *Nature*, 323(6088):533–536.

S.R. Safavian and D. Landgrebe. 1991. A survey of decision tree classifier methodology. *IEEE Transactions on Systems, Man, and Cybernetics*, 21(3):660–674.

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.

Arjun Singh, Neha Patel, and Rui Zhang. 2024. Multimodal detection of cognitive distortions in patient–doctor conversations. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*.

Kristina Toutanova, Dan Klein, Christopher D. Manning, and Yoram Singer. 2003. Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of NAACL-HLT*, pages 173–180.

Ruiyi Wang, Stephanie Milani, Jamie Chiu, Jiayin Zhi, Shaun Eack, Travis Labrum, Samuel Murphy, Nev Jones, Kate Hardy, Hong Shen, et al. 2024. Patient-: 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.

Thomas Wolf et al. 2020. Transformers: State-of-the-art natural language processing.