Generalizable Cross-Lingual Cognitive Distortion Detection with Standardized Annotations and Multi-Task Learning

Hongzhi Qi[1]†, Nan Bai[2]†, Jianqiang Li[1], Wei Zhai[1], Qing Zhao[1], Qi Gao[3], Bing Xiang Yang[2], Guanghui Fu[4] *

[1]College of Computer Science, Beijing University of Technology, Beijing, China
[2]School of Nursing, Wuhan University, Wuhan, China
[3]Beijing-Dublin International College, Beijing, China
[4]Sorbonne Université, ICM, CNRS, Inria, Inserm, AP-HP, Hôpital de la Pitié-Salpêtrière, Paris, France

Abstract

Cognitive distortion is a critical issue in psychology, with most existing studies based on Burns' cognitive distortion theory. However, differences in annotation standards lead to variations in building analysis tools, resulting in inconsistent analyses and limiting the generalizability of findings, especially in large-scale and cross-linguistic contexts. To address this issue, we collected all publicly available datasets (four in total) and conducted a series of experiments to evaluate the generalizability of various cross-linguistic models. The results indicate that models exhibit significant performance differences across datasets, highlighting the generalization problem. To mitigate this issue, we propose two solutions. First, we propose a multi-task learning model based on teacher student architecture solution, which demonstrates improved generalization performance in our experiments. Second, we introduce a new dataset (~5,000 samples) derived from re-annotating existing open datasets to ensure standardized alignment. The annotation process we provided is interpretable and grounded in psychological principles. Based on this, we constructed large language models with cognitive reasoning chains, enhancing both generalizability and interpretability. This study identifies the generalization challenge in cognitive distortion research, and our experiments show that the proposed solutions significantly improve model performance. The dataset and code are publicly available at: https://github.com/HongzhiQ/CrossLinCD.

1 Introduction

With the rapid pace of modern life and the increase in stress levels, psychological health issues have become increasingly prevalent. According to the World Health Organization (WHO), one in eight people globally suffers from mental disorders (Cuijpers et al., 2023). Among them, depression is a major contributor to the global burden of diseases. Individuals with major depressive disorder (MDD) may experience profound emotional distress and, in severe cases, self-harm or suicide (Thapar et al., 2022; Ribeiro et al., 2018). Thus, early identification of depressive traits is essential for targeted interventions and suicide prevention (Orsolini et al., 2020). Cognitive factors play a central role in the onset and progression of depression (Beck, 2008). Beck's theory posits that maladaptive cognitive structures lead to negative biases, causing individuals to focus on adverse experiences, interpret events pessimistically, and engage in rumination, suppressing positive memories and exacerbated depressive symptoms (Clark and Beck, 2010; Beck, 2008). Detecting cognitive distortions early is, therefore, critical for effective prevention strategies. Using artificial intelligence (AI) for large-scale data analysis offers a promising approach to identify such distortions early and provide timely psychological support (Graham et al., 2019).

Despite these advances, developing tools for cognitive distortion recognition still faces several major challenges. First, the scarcity of publicly available datasets limits model training and evaluation (Shickel et al., 2020; Wang et al., 2023b; Elsharawi and El Bolock, 2024). Second, while existing datasets are based on the same theoretical foundations (Burns' cognitive distortion theory (Burns, 1981)), annotation discrepancies exist due to subjective differences in human labeling, leading to poor model generalization across datasets. For example, in Chinese datasets, Qi et al. (2023) use a multi-label classification with twelve categories, while Wang et al. (2023a) adopt a multi-class approach with seven. In contrast, Sharma et al. (2023) define fourteen categories for an English dataset. These issues hinder the scalability and reliability of cognitive distortion detection models, especially in multilingual contexts.

To address these challenges, we collected all

† These authors contributed equally to this work.
* Corresponding author: guanghui.fu@icm-institute.org

Findings of the Association for Computational Linguistics: ACL 2025, pages 16041–16051

publicly available cognitive distortion datasets and conducted a series of cross-lingual experiments. Our findings reveal substantial performance inconsistencies across datasets, highlighting the need for improved generalization. To mitigate this issue, we proposed two solutions. First, we developed a multi-task learning framework that enables models to capture language-invariant cognitive distortion features, improving cross-dataset performance. Second, we propose a new dataset of 5,702 samples by reannotating existing open datasets for standardized alignment. We provide a psychological principled annotation process to support the training of large language models with reasoning chains. Experimental results demonstrate that our proposed dataset leads to a significant improvement in model performance.

This study underscores the challenge of generalization in cognitive distortion recognition and presents two feasible solutions. By releasing our datasets, we aim to facilitate future research in this domain and contribute to the development of more robust and interpretable cognitive distortion detection models.

2 Related work

Cognitive distortion research is vital for understanding flawed thinking patterns that contribute to mental health issues (Singh et al., 2024). Numerous studies have focused on utilizing AI methods to identify cognitive distortions within text, such as the research by (Elsharawi and El Bolock, 2024; Shickel et al., 2020; Wang et al., 2023b; Tauscher et al., 2023). However, these studies have not made their cognitive distortion datasets publicly available, thereby limiting further research. Among the existing studies, only four datasets have been made public (Qi et al., 2023; Wang et al., 2023a; Sharma et al., 2023; Shreevastava and Foltz, 2021), and these are insufficient for comprehensive research. Moreover, some of the available datasets are relatively small, such as the study by Sharma et al.(Sharma et al., 2023) with only 698 samples.

Differences in annotation principles lead to variations in task definitions and standards. For example, in Chinese, the dataset proposed by Qi et al. (2023) differs from that of Wang et al. (2023a): the former is a twelve-category multi-label classification problem, while the latter is a seven-category multi-class classification problem. In English, the problem still exist. For example, Sharma et al. (2023) define it with fourteen categories, while Shreevastava and Foltz (2021) use ten categories. Regarding performance, Tauscher et al. (2023) proposed a model and evaluated it on two datasets, observing a significant performance gap (Macro F1-score: 0.68 on the CrowdDist dataset and 0.23 on MH-C). The gap becomes even larger when analyzing across languages, making it difficult to develop a single, generalizable tool that performs well in both settings. Although these datasets come from different sources, all these studies (Qi et al., 2023; Wang et al., 2023a; Sharma et al., 2023; Shreevastava and Foltz, 2021) are based on the same psychological theory. Therefore, we believe they share a common underlying concept for identifying cognitive distortions.

In this study, we propose two solutions to address these issues. The first involves a multi-task teacher-student framework, while the second focuses on reannotating data for standardized alignment. Experimental results show that our solutions significantly improve generalization in this task. We acknowledge the diversity of annotation principles and leave this issue as an open topic while providing our proposed solutions.

3 Dataset

We collected all publicly available datasets related to cognitive distortions to evaluate generalization across datasets. Then, we invited annotators to label them, creating a standardized and unified dataset to enhance generalization. The dataset distribution and characteristic can be seen in Table 1.

3.1 Public datasets

- **SocialCD-3k (Qi et al., 2023) (D_s):** A multi-label classification Chinese dataset containing a total of 3,407 data samples and 12 labels. The dataset was sourced from the comment data of "Zoufan" (ZouFan, 2023) on the Weibo social media platform.

- **C2D2 Dataset (Wang et al., 2023a) (D_c):** A single-label classification Chinese dataset consisting of 7,500 data samples, each with 7 labels. To simplify annotation, this study treats the task as multi-class classification, assigning only the dominant cognitive distortion when multiple distortions are present. This dataset is the only one treated as a multi-class classification problem, making direct evaluation impossible. Therefore, we used it as

Table 1: The data distribution of the experimental datasets. $\overline{\mathbf{L}}$ represents the average number of labels per sample, and $\overline{\mathbf{C}}$ represents the average number of characters per sample. N_{train}, N_{val}, N_{test} represent the number of samples in the training, validation, and test sets, respectively.

Categories	D_s	D_c	D_r	D_{t1}	D_{t2}	D_a
Emotional reasoning	16	751	48	169	169	91
Overgeneralization	141	894	115	277	277	380
Labeling	1961	721	104	203	203	2186
Mind reading	121	1003	81	295	295	416
Fortune-telling	652	682	85	210	210	864
All-or-nothing thinking	77	690	97	126	126	166
Should statements	84		53	135	135	171
Magnification	321		110	245	245	417
Personalization	188	709	80	202	202	329
Mental filter	378			151	151	487
Disqualifying the positive	27		59			46
Blaming	27		39			41
Negative feeling or emotion			102			
Comparing and despairing			19			
Comparing			8			
Language	ZH	ZH	EN	EN	EN	ZH+EN
$N_{categories}$	12	7	14	10	10	12
Paradigm	ML	MC	ML	ML	ML	ML
$\overline{\mathbf{L}}$	1.71	1	1.43	1.26	0.80	0.98
$\overline{\mathbf{C}}$	42.56	29.68	91.34	188.70	869.81	89.46
N_{train}	2043	4501	418	958	1518	3420
N_{val}	682	1500	140	319	506	1141
N_{test}	682	1500	140	320	506	1141
Total	3407	7500	698	1597	2530	5702

a corpus to conduct experiments based on a teacher-student architecture.

- **Cognitive Reframing Dataset (Sharma et al., 2023) (D_r):** A multi-label classification English dataset containing 698 data samples, each with 14 labels.

- **Therapist Dataset (Shreevastava and Foltz, 2021) (D_t):** A multi-label classification English dataset, containing 10 types of cognitive distortions. In this study, due to differences in data volume, we denote them as D_{t1} with 1,597 samples and D_{t2} with 2,530 samples.

3.2 Standardized alignment dataset (D_a)

To improve the generalization of cognitive distortions across language models, we re-annotated three public datasets (D_s, D_r and D_{t1}) according to Burns' theory and provided a detailed annotation process, denoted as D_a. The annotation process can be seen in Figure 1. Since the Therapist-2K dataset includes more non-cognitively distorted data compared to Therapist-1K, we did not re-annotate it. The data distribution for D_a is shown in Table 1 and Figure 2.

Cognitive distortion annotation In the dataset construction's early stages, we created an annotation guide with descriptions of 12 cognitive distortion labels, examples, and counteracting strategies. Two English-fluent psychology graduate students were recruited as annotators. We conducted two rounds of annotation training: first, explaining the theories behind cognitive distortions and reviewing the guide; second, using sample cases to check understanding and accuracy. To assess inter-annotator reliability, we randomly selected 50 samples for joint annotation. Once Cohen's Kappa reached 85.51%, we continued with the full annotation. For ambiguous cases, domain experts with more than 10 years experience and annotators held discussions to finalize the labels. This resulted in the standardized dataset D_a with 5,702 samples, categorized into 12 cognitive distortion labels.

Reasoning chain annotation We constructed cognitive chains to explain label annotations in the aligned dataset. Domain experts first selected a typical sample for each label and provided detailed explanations of the cognitive distortion and reasoning. By employing prompt engineering, we leveraged the GPT-4 API (OpenAI, 2023) twice for each sample to generate cognitive reasoning chain information for the remaining samples. The domain experts then selected the highest-quality reasoning chain for each sample. Finally, the domain experts examined each reasoning chain until it met the required quality standards. The final cognitive reasoning chain, denoted as D_a^c.

4 Experimental methods

This study does not focus on algorithmic innovation but rather on addressing the generalization problem across datasets and languages in the cognitive distortion classification task. Therefore, for deep learning models, our experimental approach uses XLM (Conneau et al., 2020) as the baseline due to its strong cross-lingual ability. For LLMs, we selected several popular open-source models.

4.1 Cross-linguistic baseline: XLM

We selected XLM-RoBERTa (Conneau et al., 2020), a multilingual analysis model, as our baseline. It is a variant of RoBERTa (Liu et al., 2019), pre-trained on 2.5 TB of CommonCrawl data[1] from 100 languages. Its key features include extended training steps, dynamic masking, and unigram SentencePiece tokenization, enabling consistent cross-language processing.

[1]https://commoncrawl.org/

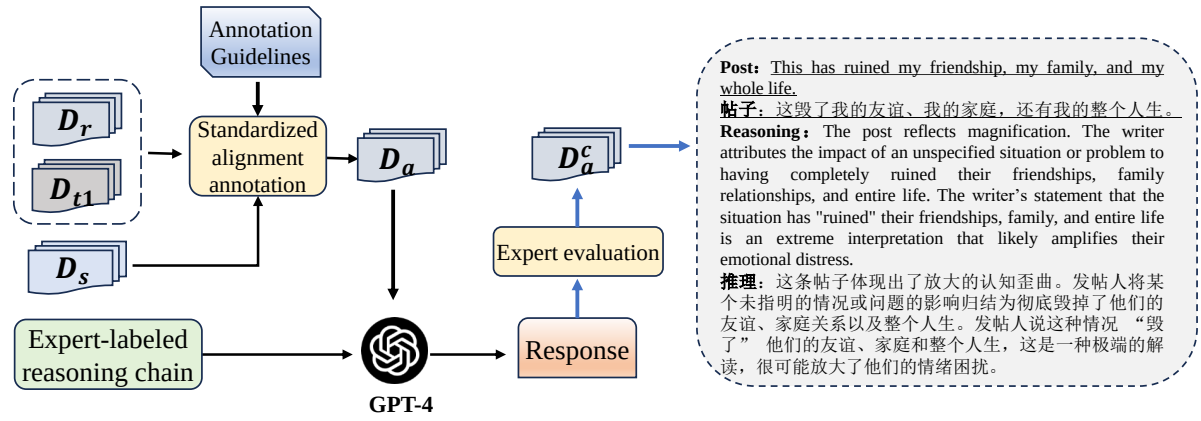

Figure 1: The annotation process of standardized alignment dataset.

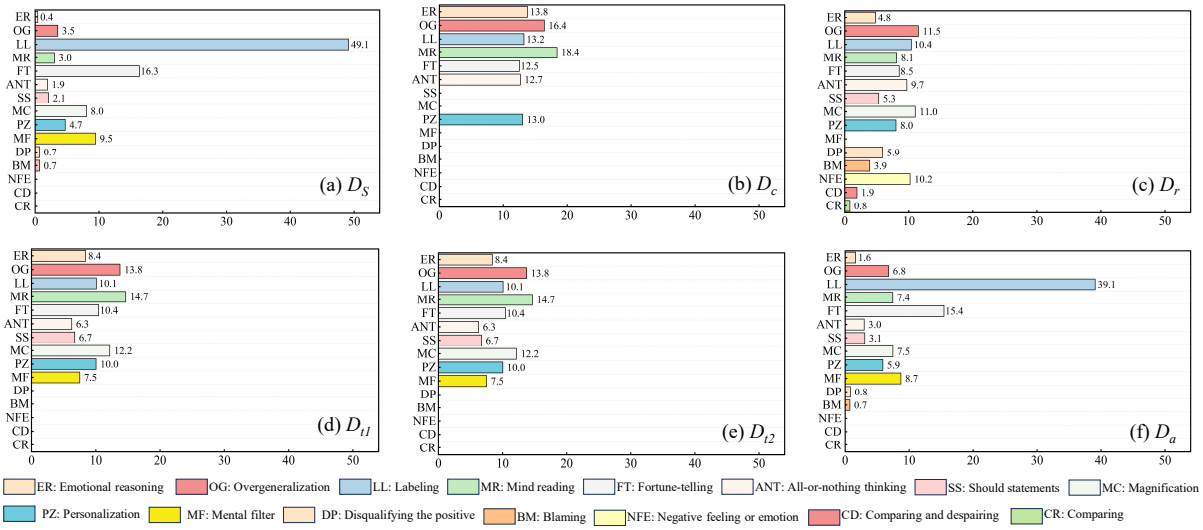

Figure 2: The distribution of label (in %) for each cognitive distortion dataset.

4.2 Large language model

LLaMA3-8B-Chinese-chat (Wang et al., 2024) : A Chinese fine-tuned LLaMA, fully fine-tuned on a mixed Chinese-English dataset of approximately 100K preference pairs, excels in both Chinese and English tasks.

GLM4 (GLM et al., 2024) : A bidirectional Transformer model with RMSNorm, SwiGLU, and extended Rotational Positional Embedding (RoPE). It uses grouped query attention for efficient KV cache use and improved autoregressive masking. Pre-trained on 10 trillion tokens with 9 billion parameters, it supports 128K token context and aligns with human preferences. According to the SuperCLUE leaderboard (Xu et al., 2023), as of December 2024, this model ranks among the top two among the 10B-level models[2].

[2]https://www.superclueai.com/

Qwen2.5-7B-Instruct (Yang et al., 2024) : Pre-trained on 7 trillion tokens with 7 billion parameters, it is fine-tuned with human feedback to improve instruction-following. Similarly, this model ranks among the top two in the 10B-level models on the SuperCLUE leaderboard, so it was selected for the experiment.

5 Experiments design

5.1 Deep learning training

5.1.1 Single task learning

To evaluate the cross-lingual generalization ability of the cognitive distortion model across different languages and datasets, we first trained four separate models using four multi-label classification datasets. Finally, we evaluated the generalization performance of the trained models by testing them on all datasets, including unseen ones.

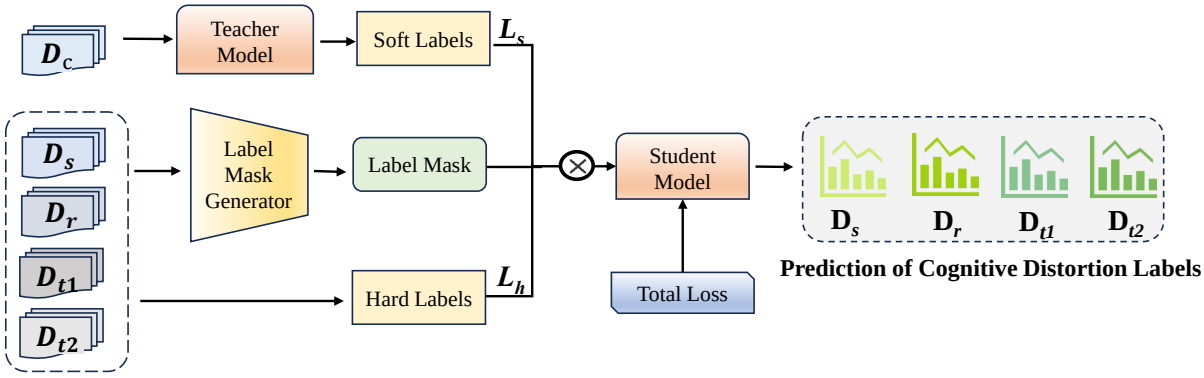

Figure 3: Training process of teacher-student model in multi-task learning.

In another words, the model M_{D_n} represents the baseline XLM model trained on dataset D_n and is evaluated across all four datasets.

5.1.2 Multi task learning

To enhance the model's generalization ability, we explore the multi-task learning paradigm. We train a multi-task model based on the XLM-RoBERTa baseline MT and evaluate it across all four datasets. In this framework, the model first encodes the input text using a shared encoder and then performs task-specific predictions through distinct classification heads. Since the model is exposed to patterns from all datasets during training, it learns to adapt to diverse annotation schemes. Each classification head is a fully connected layer, tailored to the number of labels for the corresponding task. The model is trained with task-specific labels, and during each batch update, losses for different tasks are calculated independently and optimized separately.

5.1.3 Teacher student training strategy

Given the limited availability of datasets in this field, we also experimented with a teacher-student(ts) training strategy to enhance generalization. We constructed teacher and student modules, using D_c to train the teacher module, which then provided soft labels for the student module. The student module was trained on both the target task data and the corresponding soft-label data pairs, resulting in hard loss (L_h) and soft loss (L_s). The final loss $L\mathbf{L}$ is a weighted combination of the two loss function (α=0.8 in the experiment), defined as:

$$\mathbf{L} = \alpha \cdot L_h + (1 - \alpha) \cdot L_s \quad (1)$$

This experiment also employed the XLM-RoBERTa model (Conneau et al., 2020) as the baseline, and the training process can be seen in Figure 3.

5.2 Large language model fine tuning

5.2.1 Multi task fine-tuning

In this experiment, we fine-tuned three large language models—LLaMA3-8B, Qwen2.5-7B, and GLM-4-9B—on all experimental datasets using a multi-task learning approach, as shown below:

$$M'_{\text{LLM}} = F\left(M_{\text{LLM}}, \sum D_n\right) \quad (2)$$

where M_{LLM} represents the original pre-trained language model, D_n denotes the dataset, and F is the fine-tuning function that integrates multiple datasets into the model training process. This approach highlights the flexibility of LLMs, as they impose no strict constraints on input formats, making them adaptable to various tasks.

5.2.2 Cognitive reasoning chain instructions fine-tuning

We performed instruction fine-tuning based on our proposed cognitive reasoning chain D_a^c. We selected LLaMA3-8B for reasoning chain instruction fine-tuning due to its strong bilingual capabilities in Chinese and English, resulting in LLaMA3-8B-I, which can generate explanations alongside predictions.

5.3 Implementation details

The dataset was split into training, validation, and test sets in a 6:2:2 ratio with a random seed of 42.

Deep learning model experiments The XLM-RoBERTa-large model was trained using the AdamW optimizer with a learning rate of 1×10^{-5}, batch size of 16, a threshold of 0.25, and a maximum text length of 200. The classification head was set to 15, corresponding to the union of all categories (including the "non" category). We adopts a label masking strategy based on label existence.

When processing data, a 15-dimensional label mask is created, where existing labels are set to 1 and non-existent ones to 0. During training and evaluation, the mask filters out non-existent labels, ensuring loss and metrics are calculated only for valid labels.

LLM experiments For the implementation of LLMs, we employed LoRA (Low-Rank Adaptation) (Hu et al., 2022) for parameter-efficient tuning. During training, we utilized mixed-precision training, gradient accumulation, and cosine learning rate scheduling as optimization strategies to reduce computational resource consumption and improve training stability and efficiency. For all the large models, we set the batch size to 8, the number of epochs to 5, and the learning rate to 1e-5, and tested the model that performed best on the validation set. The LLaMA3-8B-Chinese model was fine-tuned on the standardized alignment dataset for a maximum text length of 1500 tokens.

To evaluate the model performance, we computed the metrics using micro-averaging and reported the F1-score as the evaluation metric.

6 Results

Cross-dataset generalization performance Our study investigates the cross-dataset transferability of models trained on different cognitive distortion datasets. The results can be seen in Table 2.

The result reveal that while models generally perform well on their respective training datasets, their performance drops significantly when applied to other datasets. This suggests that the models lack robustness in cross-dataset generalization, likely due to inconsistencies in annotation standards across different datasets. For instance, the model trained on D_r achieves 56.17% accuracy when evaluated on the same dataset but performs significantly worse on D_{t1} (31.33%), D_{t2} (13.99%), and D_s (26.33%). Similar trends are observed across other models, indicating that the learned representations are highly dataset-specific and do not generalize well to other data distributions. This highlights the challenge posed by non-uniform labeling criteria, which introduce variations in data interpretation and hinder knowledge transfer. We also incorporated a teacher-student training framework to enhance generalizability. While some improvements are observed, such as a slight increase in average performance across datasets (from 31.96% to 33.51% for M_{D_r} and from 38.69% to 42.18% for M_{D_s}), the overall gains remain limited. Notably, the TS approach improves performance on certain datasets (e.g., $M^{ts}_{D_s}$ achieves 75.28% on D_s compared to 74.52% in the baseline), but fails to provide substantial improvements in cross-dataset settings, with performance remaining relatively low. These findings suggest that the core issue lies in the lack of standardized labeling across datasets, which restricts model transferability. While teacher-student training offers marginal benefits, it does not fully mitigate the inconsistencies caused by dataset-specific annotation schemes. This underscores the necessity of developing a unified annotation framework or multi-task learning strategies to enhance model generalization across cognitive distortion datasets.

Multi-task learning paradigms for generalization To improve cross-dataset transferability, we explore a multi-task learning approach, where models are trained jointly on multiple datasets with diverse annotation standards. This strategy aims to assess whether exposure to a broader range of data distributions can enhance generalization performance on the target datasets. As shown in Table 3, we evaluate the baseline XLM model in a multi-task learning setting (MT), its variant with a teacher-student training framework (MT^{ts}), and compare them against large language models (LLMs), including LLaMA3-8B, GLM-4-9B, and Qwen2.5-7B.

The results indicate that multi-task learning improves overall generalization. The average performance of the baseline XLM model in the multi-task setting (MT) reaches 55.84%, which surpasses the single-task training results reported in Table 3. Further improvements are observed with the teacher-student framework (MT^{ts}), achieving an average accuracy of 58.92%, demonstrating its potential to enhance model robustness. Notably, these multi-task models consistently outperform the evaluated LLMs, suggesting that task-specific supervised learning remains critical for cognitive distortion classification.

Despite these gains, multi-task learning does not surpass models trained exclusively on individual datasets. For instance, while MT^{ts} achieves 73.94% on D_s, this remains lower than the model trained specifically on D_s in a single-task setting. Similar trends are observed for other datasets, indicating a trade-off: multi-task learning improves generalization at the cost of reduced specialization. Additionally, LLMs underperform in this setting,

Table 2: Performance (F1-score) under single-task learning paradigms. Here, D_s denotes the socialCD-3k dataset (Qi et al., 2023), D_c refers to the C2D2 dataset (Wang et al., 2023a), and D_r represents the cognitive reframing dataset (Sharma et al., 2023). Additionally, D_{t1} and D_{t2} correspond to datasets of different sizes derived from the therapist dataset (Shreevastava and Foltz, 2021), while D_c indicates the re-annotated dataset. The notation M_{D_n} represents the baseline XLM model trained on dataset D_n, whereas $M^{ts}_{D_n}$ denotes the same model trained using a teacher-student learning framework.

Data/Model	M_{D_s}	$M^{ts}_{D_s}$	M_{D_r}	$M^{ts}_{D_r}$	$M_{D_{t1}}$	$M^{ts}_{D_{t1}}$	$M_{D_{t2}}$	$M^{ts}_{D_{t2}}$
D_s	74.52%	75.28%	26.33%	32.10%	32.45%	45.80%	29.96%	39.42%
D_r	23.66%	33.63%	56.17%	55.93%	21.93%	33.94%	14.23%	24.30%
D_{t1}	35.68%	38.75%	31.33%	31.25%	50.23%	50.26%	58.68%	50.53%
D_{t2}	20.91%	21.06%	13.99%	14.74%	37.75%	36.25%	24.01%	25.41%
AVG	38.69%	42.18%	31.96%	33.51%	35.59%	41.56%	31.72%	34.92%

Table 3: Performance (F1-score) under multi-task learning paradigms. Here, D_s denotes the socialCD-3k dataset (Qi et al., 2023), D_c refers to the C2D2 dataset (Wang et al., 2023a), and D_r represents the cognitive reframing dataset (Sharma et al., 2023). Additionally, D_{t1} and D_{t2} correspond to datasets of different sizes derived from the therapist dataset (Shreevastava and Foltz, 2021), while D_c indicates the re-annotated dataset. The notation MT represents the baseline XLM model trained on four datasets in a multi-task learning setting, while MT^{ts} denotes the same model trained with a teacher-student learning framework.

Data/Model	MT	MT^{ts}	LLaMA3-8B	GLM-4-9B	Qwen2.5-7B
D_s	72.00%	73.94%	54.66%	49.30%	46.15%
D_r	49.87%	55.58%	45.35%	38.55%	39.35%
D_{t1}	61.52%	64.42%	38.14%	28.38%	34.06%
D_{t2}	39.95%	41.75%	20.95%	18.68%	19.52%
AVG	55.84%	58.92%	39.78%	33.73%	34.77%

due to the task's difficulty.

These findings highlight that multi-task learning effectively enhances cross-dataset generalization by exposing the model to diverse annotation standards. However, it does not fully resolve the inconsistencies introduced by differing labeling schemes.

Ablation study on dataset composition As the teacher-student guided multi-task learning model MT^{ts} demonstrated the best performance in Table 3, we further investigate how the number of training datasets affects model performance under this architecture. To this end, we conducted an ablation study on dataset quantity in a multi-task learning setup. Specifically, we incrementally added datasets during training and evaluated the classification performance across source and target domains. The results are presented in Table 4.

We observe that the performance on D_r, D_{t1}, and D_{t2} consistently improves as more datasets are incorporated into the multi-task training process, indicating effective positive transfer and enhanced generalization in more challenging or underrepresented domains. Meanwhile, the performance on D_s remains relatively stable across different dataset configurations, suggesting that the core task signals are preserved even as additional training domains are introduced. These findings highlight a typical generalization–specialization trade-off in multi-task learning. Adding more datasets brings potential for improved generalization, especially on diverse or low-resource domains, but also risks performance drops due to domain mismatches.

Standardlized alignment for generalization To further improve cross-dataset transferability, we propose a second solution: expert reannotation to establish a standardized dataset. This approach aims to address the inconsistencies in existing datasets by aligning labeling criteria through expert judgment. The results on the standardized alignment dataset are presented in Table 5. Given the bilingual nature of the dataset, we analyze model performance separately for English (D^E_a) and Chinese (D^Z_a).

The baseline XLM model trained on the full dataset (M_{D_a}) achieves the highest overall performance, with 66.92% F1-score across both languages. When evaluated separately, the model trained on Chinese ($M_{D^Z_a}$) achieves 75.20% F1-

Table 4: Ablation study on the performance (F1-score) impact of dataset quantity under the teacher-student multi-task learning framework. Here, the notation M denotes the baseline XLM model trained on different datasets within this framework.

Data/Model	$M^{ts}_{D_s}$	$M^{ts}_{D_s+D_r}$	$M^{ts}_{D_s+D_r+D_{t1}}$	$M^{ts}_{D_s+D_r+D_{t1}+D_{t2}}$
D_s	75.28%	73.84%	73.90%	73.94%
D_r	33.63%	52.53%	53.33%	55.58%
D_{t1}	38.75%	39.47%	51.70%	64.42%
D_{t2}	21.06%	22.51%	38.94%	41.75%
AVG	42.18%	47.09%	54.47%	58.92%

Table 5: Performance (F1-score) on the standardized alignment dataset D_a. To better analyze model performance across languages, we divide the dataset into D^E_a for English and D^Z_a for Chinese. The notation M_{D_n} represents the baseline XLM model trained on dataset D_n. LLMs without '-I' are finetuned by D_a and the LLMs with 'I' is fine-tuned by D^c_a which include both data-label pair and cognitive reasoning chain.

	$M_{D^E_a}$	$M_{D^Z_a}$	M_{D_a}	LLaMA3-8B	GLM-4-9B	Qwen2.5-7B	LLaMA3-8B-I
D^E_a	40.34%	38.54%	46.42%	29.00%	18.74%	18.92%	30.97%
D^Z_a	51.30%	75.20%	75.00%	49.72%	51.27%	48.32%	54.37%
D_a	46.80%	62.65%	66.92%	44.64%	45.90%	43.16%	48.41%

score on D^Z_a, significantly outperforming other models, indicating that the reannotated Chinese dataset is well-structured and internally consistent.

We also evaluate fine-tuned large language models (LLMs), including LLaMA3-8B, GLM-4-9B, Qwen2.5-7B, and LLaMA3-8B-I. While these models exhibit lower accuracy than the deep learning baselines, they bring an additional advantage: interpretability, offering more transparent decision-making processes. Although its overall performance (48.41%) is behind that of traditional deep learning models, its ability to generate interpretable results makes it a valuable complementary approach. Comparing LLaMA3-8B-I, trained with the cognitive reasoning chain D^c_a, to LLaMA3-8B, trained on D_a, we observe the benefits of integrating the reasoning process, with a 3.77% point improvement in F1-score.

These findings suggest that standardized aligned data annotation can substantially enhance dataset consistency and improve model generalization. Furthermore, while deep learning models still achieve higher classification accuracy, LLMs contribute meaningful interpretability, which is crucial for cognitive distortion analysis. Future research could explore hybrid strategies that integrate high-accuracy deep learning models with LLM-based explainability mechanisms to achieve both strong performance and interpretability.

7 Discussion

In the task of cognitive distortion recognition, the challenge of generalization arises due to significant annotation differences across datasets, despite being grounded in the same theoretical framework. These inconsistencies limit model transfer and require strategies to enhance generalization. To address this, we explored multiple approaches, including multi-task learning, standardized annotation alignment.

Multi-task learning was investigated as a means to enhance generalization by enabling the model to learn from multiple datasets simultaneously. Given that different datasets may emphasize distinct aspects of cognitive distortion, training a shared model across multiple tasks allows for a more comprehensive representation. This approach mitigates dataset-specific biases and encourages the model to learn transferable features that improve robustness across varied data distributions.

To further reduce inconsistencies, we introduced a standardized dataset by aligning annotation criteria across different sources. Standard alignment plays a crucial role in reducing variability introduced by subjective annotation differences. By defining a unified labeling scheme, we ensured that the model is trained on a consistent dataset, minimizing annotation-induced discrepancies. However, while standardization improves uniformity, excessive alignment may inadvertently obscure dataset-specific nuances, necessitating a careful

balance between standardization and preserving valuable diversity in the data. To preserve such nuances, one promising direction is to retain original annotation metadata during training—such as label variants or annotator notes—as auxiliary signals. Additionally, future work could explore multi-view contrastive learning frameworks, where models simultaneously learn from both the standardized and the original label structures. Such designs may offer a compromise between generalizability and granularity, enabling better cultural and contextual sensitivity in cross-lingual cognitive distortion analysis.

Beyond data alignment, we explored interpretable LLMs to enhance generalization by increasing transparency in decision-making. Interpretability allows models to provide explicit reasoning for their predictions, which not only aids in model validation but also increases trust and usability. Moreover, an interpretable LLM can help uncover potential biases in training data and suggest corrective measures to refine learning strategies. However, despite the benefits of interpretability, our experiments have shown that the current performance is suboptimal. It highlights the need for future research to explore ways to enhance model performance without sacrificing transparency.

Our findings indicate that generalization in cognitive distortion recognition requires a multifaceted approach, combining structured learning strategies, data standardization, and model interpretability. While each of these strategies contributes to mitigating dataset-specific limitations, their combined effect is likely to yield the most robust improvements in generalization. Future research can further investigate the interactions between these methods and explore additional domain adaptation techniques to refine generalization across diverse datasets. In addition, it is crucial to consider ethical and practical risks when handling sensitive mental health data. Although our study relies on publicly available and anonymized sources, we acknowledge the possibility of re-identification and unintended harm. Thus, we advocate for stricter anonymization protocols, human-in-the-loop deployment, and future research into privacy-preserving techniques and model auditing tools to ensure responsible use of such technologies.

8 Conclusion

This study explored strategies to improve generalization in cognitive distortion recognition, addressing annotation inconsistencies and model limitations. We propose a multi-task learning approach and a standardized dataset to enhance generalization. Our results demonstrate that these two solutions significantly improve generalization, and the proposed open-source dataset can further support research in this domain. Future work should focus on domain adaptation, and further improving model interpretability and robustness.

Limitations

This study has several limitations. The model's performance is constrained by the availability and consistency of bilingual cognitive distortion datasets, and cross-lingual generalization remains a challenge due to cultural and linguistic differences. While the large language model's interpretability is improved using few-shot prompting, human-level reasoning is not fully captured. Additionally, the model focuses on text-based data, limiting its applicability to multimodal mental health assessment.

Ethics statement

Ethically, we ensure data anonymity and fairness in annotation. The model is designed for research purposes, not clinical diagnosis, and potential biases remain a concern. Responsible AI development, transparency, and further evaluation are necessary to mitigate risks and enhance real-world applicability.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (Grant Nos. 72174152, 72474166), the Fundamental Research Funds for the Central Universities (Grant Nos. 2042022kf1218, 2042022kf1037), and the Young Top-notch Talent Cultivation Program of Hubei Province.

References

Aaron T Beck. 2008. The evolution of the cognitive model of depression and its neurobiological correlates. *American journal of psychiatry*, 165(8):969–977.

David D Burns. 1981. *Feeling good*. Signet Book.

David A Clark and Aaron T Beck. 2010. Cognitive theory and therapy of anxiety and depression: Convergence with neurobiological findings. *Trends in cognitive sciences*, 14(9):418–424.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

P. Cuijpers, A. Javed, and K. Bhui. 2023. The who world mental health report: a call for action. *British Journal of Psychiatry*, 222(6):227–229.

Nada Elsharawi and Alia El Bolock. 2024. C-journal: A journaling application for detecting and classifying cognitive distortions using deep-learning based on a crowd-sourced dataset. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 3224–3234.

Team GLM, :, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Dan Zhang, Diego Rojas, Guanyu Feng, Hanlin Zhao, Hanyu Lai, Hao Yu, Hongning Wang, Jiadai Sun, Jiajie Zhang, Jiale Cheng, Jiayi Gui, Jie Tang, Jing Zhang, Jingyu Sun, Juanzi Li, Lei Zhao, Lindong Wu, Lucen Zhong, Mingdao Liu, Minlie Huang, Peng Zhang, Qinkai Zheng, Rui Lu, Shuaiqi Duan, Shudan Zhang, Shulin Cao, Shuxun Yang, Weng Lam Tam, Wenyi Zhao, Xiao Liu, Xiao Xia, Xiaohan Zhang, Xiaotao Gu, Xin Lv, Xinghan Liu, Xinyi Liu, Xinyue Yang, Xixuan Song, Xunkai Zhang, Yifan An, Yifan Xu, Yilin Niu, Yuantao Yang, Yueyan Li, Yushi Bai, Yuxiao Dong, Zehan Qi, Zhaoyu Wang, Zhen Yang, Zhengxiao Du, Zhenyu Hou, and Zihan Wang. 2024. Chatglm: A family of large language models from glm-130b to glm-4 all tools.

Sarah Graham, Colin Depp, Ellen E Lee, Camille Nebeker, Xin Tu, Ho-Cheol Kim, and Dilip V Jeste. 2019. Artificial intelligence for mental health and mental illnesses: an overview. *Current psychiatry reports*, 21:1–18.

Edward Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. Lora: Low-rank adaptation of large language models. In *Proceedings of the International Conference on Learning Representations (ICLR)*. ICLR. Microsoft Corporation.

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. *arXiv preprint arXiv:1907.11692*.

OpenAI. 2023. Gpt-4 technical report. https://cdn.openai.com/papers/gpt-4.pdf. Accessed: 2023-10-01.

Laura Orsolini, Roberto Latini, Maurizio Pompili, Gianluca Serafini, Umberto Volpe, Federica Vellante, Michele Fornaro, Alessandro Valchera, Carmine Tomasetti, Silvia Fraticelli, et al. 2020. Understanding the complex of suicide in depression: from research to clinics. *Psychiatry investigation*, 17(3):207.

Hongzhi Qi, Qing Zhao, Changwei Song, Wei Zhai, Dan Luo, Shuo Liu, Yi Jing Yu, Fan Wang, Huijing Zou, Bing Xiang Yang, et al. 2023. Evaluating the efficacy of supervised learning vs large language models for identifying cognitive distortions and suicidal risks in chinese social media. *arXiv preprint arXiv:2309.03564*.

Jessica D Ribeiro, Xieyining Huang, Kathryn R Fox, and Joseph C Franklin. 2018. Depression and hopelessness as risk factors for suicide ideation, attempts and death: meta-analysis of longitudinal studies. *The British Journal of Psychiatry*, 212(5):279–286.

Ashish Sharma, Kevin Rushton, Inna Lin, David Wadden, Khendra Lucas, Adam Miner, Theresa Nguyen, and Tim Althoff. 2023. Cognitive reframing of negative thoughts through human-language model interaction. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9977–10000, Toronto, Canada. Association for Computational Linguistics.

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 Singh, Sai Vemulapalli, Mauajama Firdaus, and Asif Ekbal. 2024. Deciphering cognitive distortions in patient-doctor mental health conversations: A multimodal llm-based detection and reasoning framework. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 22546–22570.

Justin S Tauscher, Kevin Lybarger, Xiruo Ding, Ayesha Chander, William J Hudenko, Trevor Cohen, and Dror Ben-Zeev. 2023. Automated detection of cognitive distortions in text exchanges between clinicians and people with serious mental illness. *Psychiatric services*, 74(4):407–410.

Anita Thapar, Olga Eyre, Vikram Patel, and David Brent. 2022. Depression in young people. *The Lancet*, 400(10352):617–631.

Bichen Wang, Pengfei Deng, Yanyan Zhao, and Bing Qin. 2023a. C2d2 dataset: A resource for the cognitive distortion analysis and its impact on mental health. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10149–10160.

Bichen Wang, Yanyan Zhao, Xin Lu, and Bing Qin. 2023b. Cognitive distortion based explainable depression detection and analysis technologies for the adolescent internet users on social media. *Frontiers in Public Health*, 10:1045777.

S. Wang, Y. Zheng, G. Wang, S. Song, and G. Huang. 2024. Llama3-8b-chinese-chat (revision 6622a23). Accessed: 2024-12-10.

Liang Xu, Anqi Li, Lei Zhu, Hang Xue, Changtai Zhu, Kangkang Zhao, Haonan He, Xuanwei Zhang, Qiyue Kang, and Zhenzhong Lan. 2023. Superclue: A comprehensive chinese large language model benchmark. *arXiv preprint arXiv:2307.15020*.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, et al. 2024. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.

ZouFan. 2023. Sina weibo "Zoufan" comment.