Towards Consistent Detection of Cognitive Distortions: LLM-Based Annotation and Dataset-Agnostic Evaluation

Neha Sharma, Navneet Agarwal, Kairit Sirts
University of Tartu, Estonia
{neha.sharma, navneet.agarwal, kairit.sirts}@ut.ee

Abstract

Text-based automated Cognitive Distortion detection is a challenging task due to its subjective nature, with low agreement scores observed even among expert human annotators, leading to unreliable annotations. We explore the use of Large Language Models (LLMs) as consistent and reliable annotators, and propose that multiple independent LLM runs can reveal stable labeling patterns despite the inherent subjectivity of the task. Furthermore, to fairly compare models trained on datasets with different characteristics, we introduce a dataset-agnostic evaluation framework using Cohen's kappa as an effect size measure. This methodology allows for fair cross-dataset and cross-study comparisons where traditional metrics like F1 score fall short. Our results show that GPT-4 can produce consistent annotations (Fleiss's Kappa = 0.78), resulting in improved test set performance for models trained on these annotations compared to those trained on human-labeled data. While human expert verification was inconclusive on our target dataset, our findings suggest that LLMs can offer a scalable and internally consistent alternative for generating training data that supports strong downstream performance in subjective NLP tasks.

Keywords: Cognitive Distortion Detection, Large Language Models, Dataset-Agnostic Evaluation, Mental Health, Automated Annotation, LLM-as-Annotator

1. Introduction

Introduced by **Aaron T. Beck** (Beck, 1963) and expanded by **David Burns** (Burns and Good, 1980), cognitive distortions (CDs) are common patterns of biased thinking that can influence how individuals perceive and interpret reality (Beck and Alford, 2009). Research shows that CDs are often associated with mental health conditions such as depression (Joormann and Stanton, 2016), anxiety (Yazici-Çelebi and Kaya, 2022), and PTSD (Ouhmad et al., 2024). These thought patterns can often be reflected in language, making them potentially detectable through natural language analysis. With the increased availability of digital text data, especially from online forums and social media, researchers have started exploring automated CD detection (Shickel et al., 2020; Rojas-Barahona et al., 2018; Simms et al., 2017), which in turn could support the development of automated tools for cognitive reframing of negative thoughts or reappraising events causing negative emotions.

Automated CD detection requires annotated data, either for training classification models or evaluating the accuracy of LLM-based systems using in-context learning. However, CD annotation can be subjective, as annotators may perceive and interpret distortions differently depending on their understanding of the concept or the context provided. Prior works have reported low agreement between annotators, highlighting the difficulty in achieving consistency when annotating CDs (Shreevastava and Foltz, 2021; Lybarger et al., 2022; Ding et al., 2022). This subjectivity can result in annotated datasets that are noisy and potentially unreliable. Models trained on such data are forced to fit these inconsistencies, which can hinder their ability to generalize and accurately detect CDs in test data. Indeed, previous works have reported weighted F1-score in a range between 0.2–0.4 (Chen et al., 2023; Shickel et al., 2020; Lim et al., 2024), suggesting that models have considerable confusion in discriminating between different CD categories.

Recent works have explored LLMs as alternatives to expert or crowd-sourced annotators in text classification tasks, typically focusing on single predictions or comparisons with human annotations (Gilardi et al., 2023; Heseltine and Clemm von Hohenberg, 2024; Aldeen et al., 2023). However, relying on a single output from LLM as the final annotation can be problematic, as it raises questions about reliability and whether the label is trustworthy or merely a result of randomness in sampling.

Therefore, we hypothesize that multiple independent LLM runs can help surface labels that are consistently assigned across runs, pointing to a reliable feature characteristic of the text. The intuition is that while individual predictions may vary, the recurrence of certain labels suggests an underlying stability, offering a form of internal agreement that could reduce some of the variability seen in human annotations, leading to more robust labels. Building on this, we propose the following contributions:

Contribution 1: We propose an LLM-based annotation framework, where we run multiple independent annotation passes on the same text using GPT-based models, and select labels that appear consistently across nearly all runs, likely representing core CDs present in the text. We focus on

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 10866–10882
11–16 May 2026.

producing annotations that are internally reliable by identifying labels that appear consistently across multiple runs. Our view is that reliability is a necessary first step toward assessing validity, especially in subjective domains like mental health, where ground truth is often unattainable. Without consistent behavior, neither model evaluation nor human verification can meaningfully proceed.

We verify the effectiveness of our methodology through: **(i) Statistical validation of annotation reliability:** Where LLMs, especially GPT-4, with moderate temperature settings, achieve high internal consistency (Fleiss' kappa = 0.78) across multiple runs. **(ii) Performance on downstream task:** Where models trained on our LLM-generated labels consistently outperform those trained on human-annotated labels. **(iii) Verification by domain experts:** Where results were inconclusive, with no clear preference emerging between label sources. However, a low agreement between experts highlights the inherent ambiguity and subjectivity of the task and limitations in the dataset itself.

Contribution 2: Re-annotation of the dataset makes direct comparison of downstream task performance difficult across the literature. This led us to our second contribution: a dataset-agnostic evaluation methodology. We adopt the notion of the effect size calculation (commonly used in psychology and medicine), with the kappa measure, which accounts for both random chance and model performance, and provides a standardized scale to compare the performances of predictive models trained on datasets with differing characteristics. The proposed metric result shows that models trained on LLM-annotated labels consistently achieve higher scores, indicating a greater improvement over the random baseline, as compared to models trained on human-annotated labels.

In summary, we propose an LLM-based annotation method for subjective tasks, grounded in the assumption that reliable and meaningful labels will consistently emerge across multiple independent runs. We also introduce a data-agnostic method for evaluating models trained and tested on different datasets by adapting the concept of effect size, using the kappa measure to provide a performance metric normalized against chance-level baselines.

2. Related Work

Several studies have explored automated CD detection using different datasets from varying sources. However, many of these efforts suffer from low inter-annotator agreement (IAA) (Shreevastava and Foltz, 2021) or the lack of it (Simms et al., 2017; Aureus et al., 2021; Shickel et al., 2020), which raises concerns about the reliability of the annotated labels and their utility in classification tasks. Wang et al. (2023) report high IAA scores; however, the English version of the dataset is not publicly available yet. Similarly, studies based on patient-therapist text exchange (Lybarger et al., 2022; Tauscher et al., 2023; Ding et al., 2022) also report low to moderate IAA scores and limited classification performance, but the sensitive nature of the dataset makes it publicly unavailable.

LLMs like ChatGPT have emerged as scalable alternatives to human annotators, offering consistent and cost-effective labeling across various NLP tasks (He et al., 2024; Li, 2024; Zhang et al., 2023; Gilardi et al., 2023). Recent studies leveraged LLMs for CD classification task through prompting frameworks. The ERD framework (Lim et al., 2024) and DoT prompting (Chen et al., 2023) both apply multi-step reasoning to enhance CD detection. However, as they focused only on classification task based on the same publicly available but unreliable gold-standards annotations by Shreevastava and Foltz (2021) with IAA 33.7%, the core issue of annotation quality remains unexplored.

In contrast to these studies, our work addresses the foundational problem of label reliability by introducing an annotation schema utilizing LLMs aimed at producing more consistent and reliable ground truth annotations, which automatically results in better performance on down stream tasks.

3. CD Annotation with LLMs

We hypothesize that the recurrence of labels across multiple independent LLM runs reflects the model detecting stable and interpretable patterns in the input text. Rather than treating single outputs as definitive annotations, we view consistent predictions as a signal of internal model reliability, particularly valuable in subjective tasks where objective ground truth is unavailable. In this section, we first describe our data and annotation procedure, followed by a quantitative analysis of label consistency across runs. Finally, we assess the reliability of the resulting annotations using inter-run agreement metrics. Together, these steps establish the foundation for our methodology: leveraging label stability as a proxy for annotation reliability in the absence of human consensus.

3.1. Therapist Q&A dataset

In this study, we use the publicly available Therapist Q&A dataset[1], which was annotated by Shreevastava and Foltz (2021) with ten CDs, and we received the annotated dataset from the authors.

[1] https://www.kaggle.com/datasets/arnmaud/therapist-qa/data

No.	Cognitive Distortion	Description and Example
1.	Emotional Reasoning	Assuming emotions reflect reality. *Example:* "I feel worthless, so I must be a failure."
2.	Overgeneralization	Drawing broad conclusions from limited experiences. *Example:* "I failed this interview, I'll never get a job."
3.	Mental Filter	Focusing only on negative aspects. *Example:* "Everyone said my presentation was good, but one person criticized it, so it must have been terrible."
4.	Should Statements	Holding rigid expectations for oneself or others. *Example:* "I should always be calm and never get upset."
5.	All-or-Nothing Thinking	Viewing situations in extremes. *Example:* "If I'm not the best, I'm a total failure."
6.	Mind Reading	Presuming negative judgments from others. *Example:* "She didn't say hi, she must think I'm annoying."
7.	Fortune Telling	Predicting negative outcomes with certainty. *Example:* "There's no point in applying, I know I won't get accepted."
8.	Magnification	Exaggerating potential problems. *Example:* "If I mess up this report, I'll lose my job and never recover."
9.	Personalization	Taking undue responsibility for external events. *Example:* "My friend is upset, it must be something I did wrong."
10.	Labeling	Defining oneself or others by single traits. *Example:* "I missed a deadline, I'm so incompetent."

Table 1: List of Cognitive Distortions (CDs) adopted from Shreevastava and Foltz (2021). "No Distortion" is included as an 11th category when no CD is present.

The dataset consists of user-written texts referred to as USER INPUT and corresponding responses referred to as RESPONSE in this study. They annotated 2530 such pairs with 10 CDs (as shown in Table 1), which are referred to as GOLDEN LABELS in the remainder of this paper. Each USER INPUT within the dataset was annotated with one dominant distortion and, optionally, a secondary distortion. The authors measured an IAA based on about one-third of the dataset labeled by two annotators, resulting in 33.7% using the joint probability of agreement metric. More details about the dataset and reasoning behind its selection for this study can be seen in Appendix A.

3.2. LLM Annotation Procedure

We start by proposing the LLM-based annotation schema that forms the basis for our proposed hypothesis. In particular, we experimented with OpenAI's GPT-4 models via the API[2] provided through Microsoft Azure[3]. We selected GPT-4 and GPT-4o LLM models, which we pair with two temperature settings 0.5 and 0.7, to control the level of randomness in the generated outputs, allowing us to observe whether consistent labels still emerge under variable conditions (*Detailed LLM configurations are presented Appendix B and temperature selection in Appendix C*). We did not perform extensive hyperparameter tuning, since our goal is to assess the reliability of LLM annotations rather than optimize model performance. Given the sensitive nature of the dataset, Microsoft Azure's default content filter had to be disabled to allow full processing of the dataset. Furthermore, we explore two prompts to instruct the models (*Detailed prompts can be seen in Appendix D*):

Multi-Label Prompt (MLP): This prompt annotates each USER INPUT with one or more CDs considered in this study, allowing an unconstrained assignment of labels.
Ranked-Label Prompt (RLP): Following the method by Shreevastava and Foltz (2021), this prompt constrains the model to select only the most dominant CD present in the USER INPUT, with the option to add a secondary distortion if applicable.

To test our hypothesis that whether consistent patterns emerge across independent runs, we process each USER INPUT through five independent API calls per prompt type. This is done for all configurations of models and temperatures (GPT4-0.5, GPT4-0.7, GPT4o-0.5, and GPT4o-0.7), resulting in a total of 40 annotations per USER INPUT across all combinations. An example of `user input` and its CD annotations from different configurations and prompt types can be seen in Figure 1.

Despite clear instructions in prompts, we noticed several instances where LLM generated new or modified labels were not present in the predefined list of CDs. We reviewed these cases and grouped all such labels under a single category of 'Others', resulting in 12 label classes: 10 predefined CDs, No Distortion, and Others (*The list of extra labels is given in Appendix E Table 7*).

[2]https://platform.openai.com/docs/api-reference
[3]https://azure.microsoft.com/en-us

User Input. "Hello, I have a beautiful, smart, outgoing and amazing five year old little girl. Yesterday she came to me and said mom can you take me to the doctor. I ask her what was wrong and she replied: I hear voices in my ears but I do not see the people saying it. She says it happened during school during a reading circle. She thought someone called her stupid and let the teacher know. . . (continue text)"

RLP Prompt

Run	GPT4-0.5	GPT4-0.7	GPT4o-0.5	GPT4o-0.7
1	Pers.	Pers., FT	ND	ND
2	Pers.	Pers., ER	ND	ND
3	Pers.	Pers., ER	ND	ND
4	Pers.	Pers., ER	ND	ND
5	Pers.	Pers., FT	ND	ND

MLP Prompt

Run	GPT4-0.5	GPT4-0.7	GPT4o-0.5	GPT4o-0.7
1	Pers., FT	Pers.	ND	ND
2	Pers., FT	Pers.	ND	ND
3	Pers., ER	Pers.	ND	ND
4	Pers., ER	Pers.	Pers.	ND
5	Pers., FT	ND	Pers.	ND

Figure 1: Example illustrating CD annotations for one user input across different GPT configurations, prompt types, and runs. (Here: Pers. = Personalization, FT = Fortune Telling, ND = No Distortion, ER = Emotional Reasoning)

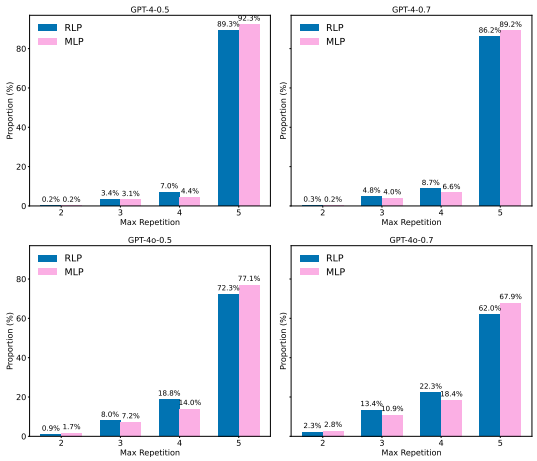

Figure 2: Distribution of maximum label repetitions across all configurations.

Configurations	RLP	MLP
GPT4-0.5	**0.78**	**0.78**
GPT4-0.7	0.73	0.71
GPT4o-0.5	0.63	0.62
GPT4o-0.7	0.52	0.54

Table 2: Average Fleiss' kappa agreement scores across model configurations and prompt types.

3.3. Label Consistency

As per our hypothesis, we explore whether repeated runs of LLMs yield more stable and consistent CD labels. To investigate this, Figure 2 plots the proportion of data points having at least one label repeated a given number of times, illustrating the extent of label consistency across runs.

Plots show most data points have at least one label repeated 5 times, indicating a high model confidence in these labels. GPT-4 at 0.5 temperature shows the highest consistency, with over 90% of data points having at least one label repeated in all runs. Across prompt types and temperatures, GPT-4 demonstrates greater consistency than GPT-4o, which tends to assign more varied sets of labels across runs. Across all configurations, at least 84% of data points have at least one label repeated four or more times. Overall, these results provide evidence in favor of our hypothesis that repeated runs can surface labels that are consistent and are less likely to be random.

3.4. Reliability

To assess the reliability of LLM-generated labels across the five independent LLM runs, we compute Fleiss' kappa (Fleiss, 1971), which is a standard statistical measure of assessing the agreement between multiple ratings of categorical values. Since our task involves multi-label annotations, we treat each label as a separate binary task (present vs absent), and calculate Fleiss' kappa scores per label to evaluate consistency. Table 2 reports Fleiss' kappa scores across configurations and prompt types averaged over the 11 CD labels considered (*Fleiss kappa calculations, along with detailed scores, are shown in Appendix F Table 8*).

The overall scores in Table 2 show moderate to substantial agreement between the different runs across all configurations and prompt types considered. In particular, GPT-4 model consistently achieved higher agreement scores than GPT-4o, which aligns with the consistent labeling patterns observed in Figure 2. These Fleiss' Kappa scores as inter-run agreements provide statistical evidence in support of our initial hypothesis that multiple LLM runs can surface stable and reliable labels within otherwise subjective tasks.

4. CD Annotation Validation

In this section, we want to evaluate the practical utility of the LLM-generated labels. The motivation of this analysis stems from the concern that training machine learning models on noisy or inconsistent labels can lead to the models trying to fit to this noise, resulting in limited generalizability and poor performance on test data. Conversely, models trained on more stable and consistent labels are expected to benefit from a stronger learning signal,

leading to better generalization. As such, we start by defining our final label selection process.

4.1. Final Label Selection

The final selection of label(s), traditionally carried out by majority voting, i.e., labels appearing in at least 3 out of 5 runs would be selected, sets a low threshold for confidence. Instead, we prioritize labels that recur in at least 4 out of 5 runs, effectively simulating a higher confidence interval, treating it as a strong indicator of model certainty, resulting in more robust and reliable labels. Data points that do not meet this threshold are assigned special categories to indicate annotation ambiguity (*see Appendix G for more details*).

This process was repeated for all the configurations considered, providing a total of 8 sets of final labels, i.e., per configuration, we have: 1) MLP LABELS, derived from the Multi-Label Prompt, 2) RLP LABELS, derived from the Rank-Label Prompt. We also include GOLDEN LABELS, taken from Shreevastava and Foltz (2021). The distribution of final labels can be seen in Appendix G Table 9.

4.2. Experimental Setup

The dataset was first split into train, development, and test sets using a 70:15:15 ratio, with stratification based on the GOLDEN LABELS. Then, we created four datasets, one for each model-temperature configuration (GPT4-0.5, GPT4-0.7, GPT4o-0.5 and GPT4o-0.7). Consistency across datasets was ensured by keeping the split assignment fixed: a USER INPUT assigned to the training set, for instance, will always belong to the training set across all datasets. After splitting, a preprocessing step was performed independently for each dataset to remove examples with ambiguous label categories (section 4.1), generated by either prompt setting. For example, if a given USER INPUT received a valid CD label from the RLP prompt but an ambiguous category label from the MLP prompt, that example was excluded from all the label settings (RLP, MLP, and GOLDEN LABELS) within the dataset. We did this filtration to avoid introducing categories not present in the GOLDEN LABELS for the sake of comparability. Any bias introduced as a result of this filtration would be systematic for all label types and should not hinder fair comparison. As a result, the number of retained examples within each split can vary slightly between datasets, resulting in four datasets of different sizes (one for each model-temperature configuration) as shown in Table 3. The train, development, and test splits maintain a 70:15:15 ratio very closely.

We use MentalRoBERTa (Ji et al., 2022) transformer-based model for three classification tasks: (1) **Binary classification** to detect the presence of CD, (2) **Multi-class classification** predicting only the dominant label of the RLP LABELS and GOLDEN LABELS, and (3) **Multi-Label classification** to detect all CDs present in a data point.

Configuration	Train	Dev	Test	Total
Original data	1771	379	380	2530
GPT4-0.5	1667	346	356	2369
GPT4-0.7	1628	344	347	2319
GPT4o-0.5	1494	306	323	2123
GPT4o-0.7	1325	291	272	1888

Table 3: Data split sizes for the original data and for the four model and temperature configurations. The splits for LLM-generated datasets are smaller than the original, because texts with ambiguous label categories have been removed.

We trained all models using standard fine-tuning procedures with five random initializations, and results were averaged. Implementation details can be found in Appendix H.

4.3. Results

Table 4 provides the weighted F1 scores for the test sets of different datasets considered within this research, averaged over 5 random model initializations (*more detailed results in appendix I*). Our results show that models trained on RLP LABELS and MLP LABELS consistently outperform the models trained on GOLDEN LABELS across different classification tasks and datasets. This aligns with our earlier expectation that more consistent labels, such as those generated through LLM, provide a stronger training signal, allowing models to generalize more effectively.

However, it is important to emphasize that these results are only indicative and not directly comparable. First, the data splits across the four datasets are not identical. Second, even within a given dataset, although the splits remains the same, the label assignments differ across GOLD LABELS, RLP LABELS, and MLP LABELS. Therefore, in the following section, we propose a methodology to compare model performances across datasets.

5. Dataset-Agnostic Evaluation Methodology

In NLP, models and methods are typically compared using benchmark datasets. For example, the annotations provided by Shreevastava and Foltz (2021) for the Therapist Q&A dataset, also used in this work, have served as a benchmark in prior studies (Chen et al., 2023; Lim et al., 2024). How-

Datasets	RLP Labels			MLP Labels			Golden Labels		
	Binary	Multiclass	Multilabel	Binary	Multiclass	Multilabel	Binary	Multiclass	Multilabel
Gpt4-0.5	0.838	0.559	0.575	0.831	N/A	**0.609**	0.768	0.384	0.311
Gpt4-0.7	**0.854**	**0.604**	0.548	0.838	N/A	0.603	0.770	0.391	0.332
Gpt4o-0.5	0.832	0.481	0.396	0.800	N/A	0.489	0.778	0.384	0.287
Gpt4o-0.7	0.809	0.476	0.428	0.829	N/A	0.474	0.813	0.395	0.338

Table 4: Weighted F1 scores (MentalRoBERTa) on corresponding test sets averaged over five initializations. N/A indicates task not applicable for the label set.

ever, when the datasets have different sizes and label distributions, direct comparison using standard evaluation metrics becomes problematic. We faced this issue in section 4.3, where model performances were not directly comparable for our four datasets of different characteristics. In this section, we propose and implement a dataset-agnostic evaluation method for comparing model performance, based on a random baseline and the concept of effect size, which is a statistical measure commonly used to aggregate results across different datasets in fields such as psychology and medicine.

5.1. Overview

In general terms, effect size is a quantitative measure of the strength or magnitude of a phenomenon (Kelley and Preacher, 2012). Effect sizes are expressed on a standardized scale, making their magnitude relatively easy to interpret. For instance, a commonly used effect size measure, Cohen's d, represents the difference between the means of two groups in units of standard deviation. The advantage of using effect size is that it enables bringing various studies conducted using different configurations and datasets to the same scale and thus make them comparable. Although effect sizes are routinely reported in psychology and medical research, the concept remains largely unexplored in NLP and machine learning (ML) more broadly. A rare exception is the paper by Henderson and Brunskill (2018), which advocates for conducting meta-analyses in ML to quantify the influence of various factors. However, the authors do not propose a concrete method for computing effect sizes.

In this paper, we propose to use the Cohen's kappa (Cohen, 1960) to quantify the effect size for predictive classification models, which accounts for both the model's performance and random chance. Simply put, Cohen's kappa provides a normalized measure representative of a model's performance relative to random chance assignment. Through this approach, rather than looking at absolute performance across datasets, models are compared based on their performance improvements over the corresponding random chance assignments, providing a fair comparison across datasets with different characteristics. The three key features to implement this method are:

1. **Random baseline ($F1_{random}$)**: Weighted F1 score calculated using random label assignments as model predictions based on the dataset's observed class distributions.
2. **Model performance ($F1_{calculated}$)**: weighted F1 scores obtained for the trained models.
3. **Normalization (κ_{F1})**: we then normalize the model's performance using the random baseline score with the kappa formula.

5.2. Random Baseline ($F1_{random}$)

To establish a realistic performance baseline, we compute the random weighted F1 score for each dataset and classification task by assigning predictions randomly according to the class distribution observed in the corresponding ground truth labels, rather than using a uniform distribution. By aligning the predicted class probabilities with the true label distribution, this approach reflects the expected performance of a model that does not learn from the data but mirrors class imbalance. These randomly assigned labels (substitute for model predictions) and ground truth labels are then used to compute the weighted F1 score ($F1_{random}$).

We start by providing a detailed derivation of the random F1 score used as a baseline in our proposed evaluation framework. For example, we consider a dataset with 3 classes (A, B, and C) with given distributions (a, b, and c) respectively such that $a + b + c = 1$. This mean the probability of selecting label A during random assignment is $P(A) = a$, with $P(B) = b$ and $P(C) = c$. This results in a confusion matrix as shown below

	Predicted A	Predicted B	Predicted C
True A	$a \times a \times N$	$a \times b \times N$	$a \times c \times N$
True B	$b \times a \times N$	$b \times b \times N$	$b \times c \times N$
True C	$c \times a \times N$	$c \times b \times N$	$c \times c \times N$

where N is the total number of data point in the dataset and the probability of a data point belonging to class A being assigned label A in random

Datasets	RLP Labels			MLP Labels			Golden Labels		
	Binary	Multiclass	Multilabel	Binary	Multiclass	Multilabel	Binary	Multiclass	Multilabel
Gpt4-0.5	0.571	0.438	**0.348**	0.574	N/A	0.283	0.515	0.257	0.138
Gpt4-0.7	**0.572**	**0.446**	0.331	**0.598**	N/A	**0.324**	0.481	0.241	0.163
Gpt4o-0.5	0.556	0.351	0.229	0.469	N/A	0.317	0.503	0.254	0.110
Gpt4o-0.7	0.551	0.309	0.279	0.531	N/A	0.310	0.475	0.250	0.169

Table 5: Kappa scores (κ_{F1}) for MentalRoBERTa across datasets and classification tasks on test sets. Higher values indicate greater improvement over the random baseline.

assignment is $a \times a$, giving true positive value for class A (TP_A) as $a \times a \times N$. The same reasoning applies to other values within the confusion matrix.

Based on this confusion matrix, precision, recall, and F1 score for class A can be calculated as:

$$\begin{aligned}\text{Precision}_A &= \frac{TP_A}{TP_A + FP_A} \\ &= \frac{a \times a \times N}{(a \times a + a \times b + a \times c) \times N} \\ &= \frac{a^2}{a(a+b+c)} = a\end{aligned}$$

$$\begin{aligned}\text{Recall}_A &= \frac{TP_A}{TP_A + FN_A} \\ &= \frac{a \times a \times N}{(a \times a + a \times b + a \times c) \times N} \\ &= \frac{a^2}{a(a+b+c)} = a\end{aligned}$$

$$\begin{aligned}F1_A &= \frac{2 \times Precision_A \times Recall_A}{Precision_A + Recall_A} \\ &= \frac{2 \times a \times a}{a + a} \\ &= a\end{aligned}$$

Similarly, calculation for class B and C provide $F1_B = b$ and $F1_C = c$. Furthermore,

$$\begin{aligned}\text{Weighted F1} &= a \times \text{F1}_A + b \times \text{F1}_B + c \times \text{F1}_C \\ &= a^2 + b^2 + c^2\end{aligned}$$

This procedure avoids sampling variance and provides a mathematical formula for a random baseline score for each dataset. To verify the accuracy and robustness of our proposed random F1 score calculation methodology, we performed empirical tests by simulating random label assignment. Empirical results closely matched the theoretical scores, with deviations well within the expected standard error, thus validating our approach (*detailed results are provided in Appendix J*).

5.3. Normalization (κ_{F1})

The normalized effect size measure is implemented with the Cohen's kappa formula, which quantifies agreement between two sets of labels while accounting for chance, given by:

$$\kappa = \frac{P_o - P_e}{1 - P_e}$$

In our case, we use the weighted F1-score to quantify the agreement between the trained model and the annotated labels. Specifically, $F1_{calculated}$ is used for the observed agreement P_o and $F1_{random}$ is for the random agreement P_e:

$$\kappa_{F1} = \frac{F1_{calculated} - F1_{random}}{1 - F1_{random}}$$

This transformation allows us to express model performance on a standardized scale where 0 indicates random chance performance, 1 indicates perfect performance, and values between 0 and 1 reflect the degree to which model predictions exceed chance. Although the level of chance performance varies across models trained on different datasets, or on the same data with different label assignments, Cohen's kappa accounts for this by incorporating dataset-specific chance agreement, making the resulting kappa scores comparable.

5.4. Results

Table 5 presents κ_{F1} scores for our four datasets generated in the previous section across three classification tasks. These scores now enable more meaningful comparison of model performance across datasets and configurations. Consider the GPT4-0.7 configuration, a multilabel model trained on RLP LABELS provides a 33.1% improvement, whereas models trained on GOLDEN LABELS only manage to improve 16.1% over the respective baselines. Overall, models trained on LLM-generated labels (RLP LABELS and MLP LABELS) consistently outperform those trained on GOLDEN LABELS across all datasets and classification tasks. This reinforces the idea that training on more consistent and reliable LLM-annotations lead to better generalization on test data, as opposed to training on noisy or inconsistent labels that hinder model performance.

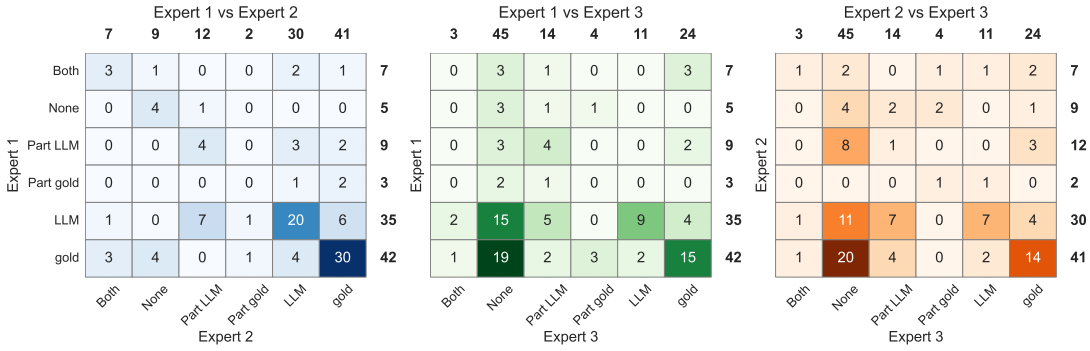

Figure 3: Pairwise confusion matrices b/w experts. (*Here, gold = GOLDEN LABELS, LLM = LLM-generated labels, part gold = Partial GOLDEN LABELS, part LLM = Partial LLM-generated labels.*) Values on the right and top of the plots represent the total number of data points within each category for respective experts.

6. Human Verification

While LLM-generated labels showed higher consistency and downstream model performance, it remains essential to assess the quality of these labels through human oversight. To this end, we conducted an expert verification experiment to validate the quality of our LLM-generated labels compared to GOLDEN LABELS.

For this process, 101 samples were randomly chosen from subset of the GPT4-0.5 dataset where LLM-generated labels and GOLDEN LABELS disagreed. GPT4-0.5 dataset was chosen based on its higher inter-run agreement (Table 2) and predictive model performance (Table 5). Three psychology experts[4] were asked to review the samples through Label Studio[5] annotation tool. They were shown the USER INPUT and a randomized, anonymized pair of LLM-generated and GOLDEN LABELS (Label1 and Label2), and asked which they agreed with: (1)Label1, (2)Label2, (3)Both, (4)None, (5)Partially Label1, or (6)Partially Label2. The experts were given the list of CDs together with their description as given by Burns (1989) as annotation guidelines (*more details in Appendix K*).

Experts' label selection and corresponding agreements can be seen in Figure 3. We observe that all three experts were nearly evenly split in their choices between the LLM-generated and Golden labels (including partial agreement), suggesting no strong preference for either label category. However, Expert 3 selected 'None' in 45% of the cases, indicating high disagreement with both label categories, in contrast to Experts 1 and 2, with 5% and 9% disagreement, respectively.

The inter-annotator agreement (Fleiss' kappa) among the three experts yielded a low score of 0.20. Furthermore, the pair-wise agreement between Expert 1 and Expert 2 was moderate (κ = 0.44), while agreement between Expert 1 and Expert 3 (κ = 0.16) and between Expert 2 and Expert 3 (κ = 0.11) were notably lower, reflecting greater disagreement. Thus, we conclude this verification process could not prove or disprove the validity of either LLM-generated or human-annotated labels, with expert opinions evenly split between the two.

7. Discussion

Given the subjectivity of the task, this study first focuses on the fundamental issue: can we produce annotations that are both internally reliable and reproducible? Our proposed method, based on repeated LLM sampling and consistency filtering, offers a systematic way to reduce annotation noise and enhance reproducibility. For tasks where ground truth is inaccessible and IAA (agreement between experts) is low, establishing reliability is not just helpful but essential for enabling any future progress on assessing or improving label quality. We further argue that reliability is a necessary first step towards assessing validity, especially in the mental health domain, where ground truth is often unattainable. Without consistent behavior, neither model evaluation nor human verification can meaningfully proceed.

To that end, we explored the use of LLMs as annotators, not in the conventional one-off prediction setting, where a single prediction is taken as final (may reflect stochastic variation rather than the model's true confidence in a label), but as a consistent pattern detector through repeated querying across multiple independent runs. Our hypothesis was that multiple LLM runs can help surface labels that are consistently assigned across runs, pointing to features that are reliable characteristics of the text. Experimental results show that LLM-generated labels have high inter-run agree-

[4]Two master-level practicing clinicians (one of them is also a co-author), and a Phd-level psychology researcher. All have at least introductory level training in CBT.

[5]https://labelstud.io/

User Text	Gold	LLM
I am writing because my boyfriend and I have a lot of problems in the one year we've been together. Six months ago we went on a break because I wanted to live with him but he didn't want to live with me. Even though I didn't want to end it, the arguments we had over the living together issue seemed to push him to the point of wanting to leave.	MR	ND
From a young woman in Bangladesh: I have been in a very physically and mentally abusive marriage for 4 years now. I tried my best to make my marriage work and meet up to my husband's and his family's expectations, but I am always being told that I am good for nothing and I should probably kill myself. I have been accused of infidelity multiple times even when I had never done anything like that. But recently, I just couldn't tolerate all that anymore.	Over	ER, AoN

Table 6: Example of USER INPUTS in data. (*Here: Gold = GOLDEN LABELS, LLM = LLM-generated labels, MR = Mind Reading, ND = No Distortion, Over = Overgeneralization, AoN = All or nothing Thinking, ER = Emotional Reasoning*)

ment (Fleiss' Kappa = 0.78 Section 3.4) as well as higher downstream task performance compared to GOLDEN LABELS (weighted F1 Section 4.3, and κ_{F1} scores Section 5.4). This gain in performance likely stems from the greater consistency of our LLM-generated labels, which offer a clearer training signal. In contrast, inconsistent labels introduce noise, hindering learning and reducing generalization, an issue seen with GOLDEN LABELS. These results not only support the use of LLMs as reliable annotators for generating consistent CD labels but also highlight the need for multiple LLM runs rather than relying on a single prediction.

In the second part of this study, we focus on the comparability problem, how to meaningfully evaluate models when datasets differ in size, label distribution, or annotation quality. Traditional evaluation procedures used in NLP and ML rely on comparing evaluation measures obtained on the same benchmark dataset. However, if the dataset composition or label assignment differs, the absolute values of the evaluation measures are no longer comparable. To overcome this limitation, we propose a kappa-based effect size measure (κ_{F1}) that normalizes model performance relative to chance, enabling dataset-agnostic and cross-study comparison. Crucially, this approach is not limited to CD detection: it generalizes to any domain where standardized baselines are unavailable or unreliable and where datasets vary in composition. Such a measure provides a principled and interpretable foundation for comparing model performance across heterogeneous datasets, promoting fairness and reproducibility in evaluation.

Although the LLM-generated labels proved to be more consistent and less noisy than the initial GOLDEN LABELS, human evaluation involving domain experts proved to be inconclusive. Upon examining examples (Table 6) where experts disagreed, we observed that many of these USER INPUTS do not explicitly state thoughts, but instead describe events, emotions, or experiences. While distorted thought patterns may underlie such accounts, the descriptions alone are often insufficient to reliably identify specific cognitive distortions. In a clinical setting, a therapist would typically first probe further into the user's underlying thoughts before identifying a specific distortion. This points towards the data source itself being a limiting factor, and we therefore suggest the development of methodologies that ensure that distorted thought patterns are adequately expressed within the text.

8. Conclusion

This study demonstrates that Large Language Models can serve as consistent and reliable annotators for inherently subjective tasks such as cognitive distortion detection. By reframing annotation as a process of repeated querying rather than single-shot prediction, we show that multiple independent LLM runs can reveal stable labeling patterns that improve inter-run agreement and yield superior downstream model performance. These findings highlight the potential of LLMs to generate internally coherent annotations that reduce noise and enhance reproducibility, an essential step toward establishing reliability in domains where ground truth is uncertain. Beyond annotation reliability, we addressed the broader challenge of model comparability across datasets with differing characteristics. The proposed kappa-based effect size measure (κ_{F1}) offers a dataset-agnostic evaluation framework that normalizes model performance relative to chance, enabling fairer and more interpretable comparisons across heterogeneous datasets and studies. Importantly, this framework generalizes beyond NLP to any field where benchmarks are unstable or direct comparison is not feasible. Finally, low agreement observed in human validation suggests that data lacking sufficiently articulated thought content may be inadequate for CD detection, calling for either more carefully curated datasets or methods that first identify and elicit the necessary cognitive elaboration before prediction.

Limitations

While our study makes useful contributions, it also has limitations. First, the annotation process relied exclusively on models from the GPT family. While these models demonstrated strong performance and consistency across runs, we did not explore the applicability or reliability of other proprietary large language models, such as Claude, Gemini, or open-source models, such as LLaMA or Mistral. As such, it remains unclear whether similar annotation quality and agreement can be achieved using non-OpenAI LLMs. We believe this represents a promising future direction for our work, involving a systematic comparison of annotation consistency and quality across different LLM families.

Second, although our methodology achieved consistency and reliability on the Therapist Q&A dataset, several dataset-specific limitations must be acknowledged. We observed that many of the USER INPUTS do not explicitly express distorted thoughts but instead describe events, emotions, or experiences. While distorted thought patterns may underlie such accounts, the descriptions alone are often insufficient to reliably identify specific distortions. In a clinical context, a therapist would typically probe further to uncover the client's underlying distortions before assigning a label. This indicates that the data source itself imposes a structural limitation: it captures surface descriptions rather than explicit cognitive patterns, which constrains interpretability and diagnostic precision. Furthermore, the dataset's source cannot be independently verified, which also raises questions about authenticity and representativeness. These factors suggest that, while our framework is methodologically sound, the dataset may not be ideal for clinically oriented or context-dependent applications. Nevertheless, this dataset was selected because it remains one of the few publicly available and comparatively higher-quality resources for cognitive distortion detection and is already in active use within the research community. Its limitations underscore a broader issue within the domain, i.e., the urgent need for contextually richer datasets that more directly capture distorted thought patterns in text.

Acknowledgments

This research was supported by the Estonian Research Council Grant PSG721 and by the Estonian Centre of Excellence in AI (EXAI). We thank the two annotators Pärtel Poopuu and Helen Uusberg.

Ethics Statement

This study makes exclusive use of the publicly available Therapist Q&A dataset and open-source language models. The dataset contains anonymized text that does not include any personally identifiable information. All analyses were conducted in accordance with the dataset's terms of use. No human subjects were directly involved, and no additional data were collected, ensuring compliance with ethical standards for secondary data research.

The broader goal of this work is to improve the reliability and transparency of automated annotation methods for subjective psychological constructs. While our findings may support the development of tools for mental health research, they are not intended for clinical diagnosis or therapeutic decision-making. We encourage responsible use of these methods within appropriate research and ethical boundaries.

Data and Code Availability

The dataset used in this study is publicly available on Kaggle[1]. The code and LLM-generated labels developed for this work can be seen at GitHub[6] .

References

Mohammed Aldeen, Joshua Luo, Ashley Lian, Venus Zheng, Allen Hong, Preethika Yetukuri, and Long Cheng. 2023. Chatgpt vs. human annotators: A comprehensive analysis of chatgpt for text annotation. In *2023 International Conference on Machine Learning and Applications (ICMLA)*, pages 602–609.

Jelly P. Aureus, Ma. Regina Justina E. Estuar, Dorothy C. Mapua, Roland P. Abao, and Anna Angeline M. Cataluña. 2021. Determining linguistic markers in cognitive distortions from covid-19 pandemic-related reddit texts. In *2021 1st International Conference in Information and Computing Research (iCORE)*, pages 56–61.

Aaron T Beck. 1963. Thinking and depression: I. idiosyncratic content and cognitive distortions. *Archives of general psychiatry*, 9(4):324–333.

Aaron T Beck and Brad A Alford. 2009. *Depression: Causes and treatment*. University of Pennsylvania Press.

David D Burns. 1989. *The feeling good handbook: Using the new mood therapy in everyday life.* William Morrow & Co.

David D Burns and MD Feeling Good. 1980. *The new mood therapy*.

[6] https://github.com/nehasharma666/llm-cognitive-distortion-detection

Zhiyu Chen, Yujie Lu, and William Wang. 2023. 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.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

Xiruo Ding, Kevin Lybarger, Justin Tauscher, and Trevor Cohen. 2022. Improving classification of infrequent cognitive distortions: Domain-specific model vs. data augmentation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Student Research Workshop*, pages 68–75, Hybrid: Seattle, Washington + Online. Association for Computational Linguistics.

Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.

Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. 2023. Chatgpt outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences*, 120(30):e2305016120.

Xingwei He, Zhenghao Lin, Yeyun Gong, A-Long Jin, Hang Zhang, Chen Lin, Jian Jiao, Siu Ming Yiu, Nan Duan, and Weizhu Chen. 2024. Annollm: Making large language models to be better crowdsourced annotators.

Peter Henderson and Emma Brunskill. 2018. Distilling information from a flood: A possibility for the use of meta-analysis and systematic review in machine learning research. *CoRR*, abs/1812.01074.

Michael Heseltine and Bernhard Clemm von Hohenberg. 2024. Large language models as a substitute for human experts in annotating political text. *Research & Politics*, 11(1):20531680241236239.

Shaoxiong Ji, Tianlin Zhang, Luna Ansari, Jie Fu, Prayag Tiwari, and Erik Cambria. 2022. MentalBERT: Publicly Available Pretrained Language Models for Mental Healthcare.

Jutta Joormann and Colin H Stanton. 2016. Examining emotion regulation in depression: A review and future directions. *Behaviour research and therapy*, 86:35–49.

Ken Kelley and Kristopher J Preacher. 2012. On effect size. *Psychological methods*, 17(2):137.

Jiyi Li. 2024. A comparative study on annotation quality of crowdsourcing and llm via label aggregation. In *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6525–6529.

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.

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

Nawal Ouhmad, Romain Deperrois, Wissam El Hage, and Nicolas Combalbert. 2024. Cognitive distortions, anxiety, and depression in individuals suffering from ptsd. *International Journal of Mental Health*, 53(4):336–352.

Lina Rojas-Barahona, Bo-Hsiang Tseng, Yinpei Dai, Clare Mansfield, Osman Ramadan, Stefan Ultes, Michael Crawford, and Milica Gasic. 2018. Deep learning for language understanding of mental health concepts derived from cognitive behavioural therapy.

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.

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.

T. Simms, C. Ramstedt, M. Rich, M. Richards, T. Martinez, and C. Giraud-Carrier. 2017. Detecting cognitive distortions through machine learning text analytics. In *2017 IEEE International Conference on Healthcare Informatics (ICHI)*, pages 508–512.

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.

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

Gülin Yazici-Çelebi and Feridun Kaya. 2022. Interpersonal cognitive distortions and anxiety: The mediating role of emotional intelligence. *International Journal of Psychology and Educational Studies*, 9(3):741–753.

Ruoyu Zhang, Yanzeng Li, Yongliang Ma, Ming Zhou, and Lei Zou. 2023. Llmaaa: Making large language models as active annotators.

Appendix

A. Therapist Q&A dataset

This section contains our reasoning behind using this dataset. When we began this study, this was the most suitable publicly available resource for cognitive distortion detection. The dataset has been widely used and cited in the field, including in prior works such as (Shreevastava and Foltz, 2021; Chen et al., 2023; Lim et al., 2024). Other available datasets in this domain (Simms et al., 2017; Aureus et al., 2021; Shickel et al., 2020) are, in our assessment, lower in quality, either due to annotation inconsistency, size, or lack of contextual richness. There are also non-public datasets from studies such as (Lybarger et al., 2022; Tauscher et al., 2023; Ding et al., 2022) which we could not access due to ethical and privacy constraints. Wang et al. (2023) produced a Chinese-language dataset, but no English version is currently available. We analyzed a small sample from the Chinese version and found that the text fragments were too short and contextually sparse to reliably detect cognitive distortions. Unfortunately, the source of the Therapist Q&A dataset could not be verified; moreover, the identity of the responders is also not confirmed. However this is irrelevant to our work, as our analysis focuses solely on the USER INPUT.

Given these limitations, we made the pragmatic decision to use the dataset in question because it was: publicly available, of comparatively higher quality than other open datasets, and already in active use by the research community. While we cannot verify the source of the data, our analyses suggest that the language in the dataset contains meaningful cognitive distortion signals to some extent and serves as a reasonable testbed for developing and evaluating annotation methodologies.

B. Large Language Models

Both LLM models considered in this research are detailed below:

- **GPT-4**: Version 0613, capable of processing up to 110K tokens/minute, with a rate limit of 60 requests/minute.
- **GPT-4o**: Version 2024-05-13, designed to handle up to 30M tokens/minute and up to 180K requests/minute.

C. Temperature

This section explains our choice of temperatures 0.5 and 0.7. We initially began our annotation process using GPT-4 with a temperature of 0.7, as it is the default setting. When switching to GPT-4o, we noticed that outputs at 0.7 were more varied across runs compared to GPT-4 at the same setting, resulting in noticeably lower consistency. Our choice of temperature 0.5 and 0.7 was based on trial and error. We want to observe: enough diversity to reflect the model's interpretive range on subjective input, while still allowing for patterns of consistency to emerge where the model repeatedly selects the same label across runs. The goal is not to maximize agreement artificially, but to explore whether the model could identify stable signals in the text under mild randomness. We intentionally avoided using a fully deterministic setting (e.g., temperature = 0), because evaluating consistency in such a deterministic system would not be logical as we can no longer evaluate whether a label appears consistently because it is strongly indicated by the input, or simply because the decoding process collapses to a single option. Such results would be trivially consistent, but not informative about the model's internal uncertainty or label robustness. Evaluation of the consistency across runs is meaningful only when the model is allowed some extent of variability in its output.

D. Annotation Prompts

Ranked-Label Prompt (RLP): As a specialized annotator in psychology with expertise in cognitive distortions, analyze the `user input` to identify any underlying cognitive distortion(s) from the specified list called COGNITIVE DISTORTIONS LIST. Your task is to determine the most dominant cognitive distortion present. If there is a secondary distortion, note it as well. In cases where multiple distortions are present, select the most dominant one as the primary label, and if necessary, include one secondary label. The response must include at most two labels. If no distortions are found, label the `user input` as **No Distortion**.

COGNITIVE DISTORTIONS LIST: [Emotional Reasoning, Overgeneralization, Mental Filter, Should Statements, All or Nothing Thinking, Mind Reading, Fortune Telling, Magnification (Catastrophizing), Personalization, Labeling]

Output Format: If cognitive distortions are identified, provide them as a comma-separated list, with the most dominant distortion listed first, followed by a secondary distortion if applicable. At most two distortions should be listed. If no distortions are found, return **No Distortion**.

Multi-Label Prompt (MLP): As a specialized annotator in psychology with expertise in cognitive distortions, analyze the `user input` to identify any underlying cognitive distortion(s) from the specified list called COGNITIVE DISTORTIONS LIST and provide results as per the **Output Format** only.

COGNITIVE DISTORTIONS LIST: [Emotional Reasoning, Overgeneralization, Mental Filter, Should Statements, All or Nothing Thinking, Mind Reading, Fortune Telling, Magnification (Catastrophizing), Personalization, Labeling]

Output Format: If any cognitive distortion(s) are identified, list them as a comma-separated list. If no distortion(s) are found, return **No Distortion**.

E. Extra Labels

Table 7 presents the extra labels annotated by LLMs.

F. Fleiss' Kappa

Fleiss' kappa is a statistical measure used to evaluate the consistency of agreement among multiple raters when they classify or categorize items. Like other kappa statistics (e.g., Cohen's kappa), it accounts for the level of agreement expected by chance. While Cohen's kappa is designed for two raters, Fleiss' kappa generalizes the approach to settings with more than two raters, making it well-suited for measuring inter-run agreement across multiple LLM outputs in our study.

Steps to Compute Fleiss' Kappa:
Step 1: Binary Label Conversion: Each label is represented in a binary format, where:

$$x_{ij} = \begin{cases} 1 & \text{if label } j \text{ is present in iteration } i, \\ 0 & \text{if label } j \text{ is absent in iteration } i. \end{cases}$$

This creates a binary matrix X of size $N \times M$, where N is the number of items and M is the number of iterations.

Step 2: Construct Contingency Tables: For each item i, calculate the number of agreements (n_{ij}) for presence (1) and absence (0) across all iterations:

$$n_{i1} = \sum_{j=1}^{M} x_{ij}, \quad n_{i0} = M - n_{i1}.$$

This can be visualized as a contingency table:

Item	n_{i1} (Presence)	n_{i0} (Absence)
1	n_{11}	n_{10}
2	n_{21}	n_{20}
$\vdots$	$\vdots$	$\vdots$
N	n_{N1}	n_{N0}

Step 3: Compute Fleiss' Kappa:
A. **Proportion of Agreement for Each Item**:

$$P_i = \frac{n_{i1}(n_{i1}-1) + n_{i0}(n_{i0}-1)}{M(M-1)}.$$

B. **Overall Agreement**:

$$P = \frac{1}{N} \sum_{i=1}^{N} P_i.$$

C. **Expected Agreement**:

$$P_e = \sum_{k=1}^{2} \left(\frac{\sum_{i=1}^{N} n_{ik}}{N \cdot M} \right)^2.$$

D. **Fleiss' Kappa**:

$$\kappa = \frac{P - P_e}{1 - P_e}.$$

Computed Fleiss' Kappa scores for 11 CDs across different configurations and prompt type can be seen in Table 8.

G. Final Label Selection

As explained in Section 4.1, we selected only those labels that appear at least four times across five runs. Data points that fail to meet this condition are considered too ambiguous to be consistently labeled by LLMs. For USER INPUTS where no label reached the threshold of four recurrences, we introduced two fallback categories to reflect the nature of the ambiguity. If none of the labels met the threshold but 'No Distortion' appeared in at least one run, we labeled the instance as 'not sure if distortion,' acknowledging the uncertainty around whether the text contains a distortion at all. In contrast, if the labels were too diverse without reaching any consensus and 'No Distortion' was absent, we marked it as 'not sure which distortion,' indicating

Config	MLP (Other Labels)	RLP (Other Labels)
GPT4-0.5	Lack of Emotional Empathy, Paranoid Thinking, Procrastination, Paranoia, Delusions, Minimization, Hallucinations, Lack of Empathy, Anxiety. **Total: 46**	Paranoia, Fantasy, Social Avoidance, Lack of Empathy, Obsession, Comparison, Minimization, Blaming, Paranoid Thinking, Guilt, Procrastination, Rumination, Social Anxiety, Secret Keeping, Delusions, Paranoia (Mind Reading), ADHD, Low Self-Esteem, Delusional Disorder, Avoidance, Trust Issues, Dependent Personality Disorder, Delusional Thinking, Unhealthy Lifestyle. **Total: 221**
GPT4-0.7	Avoidance, Comparison, Delusions, Delusions of Grandeur, Delusions of Persecution, Delusions of Reference, Hallucinations, Lack of Emotional Empathy, Lack of Empathy, Minimization, Paranoia, Paranoid Thinking, Procrastination, Social Comparison **Total: 53**	Avoidance, Blaming, Comparison, Delusion, Delusional Disorder, Delusional Thinking, Dependence, Disregard for Others' Feelings, Fantasy, Fantasy Thinking, Guilt, Hallucinations, Jumping to Conclusions, Lack of Empathy, Narcissism, Obsession, Paranoia, Paranoia (Mind Reading), Paranoid Thinking, Procrastination, Rumination, Self-harm, Social Anxiety, Social Avoidance **Total: 207**
GPT4o-0.5	Guilt Tripping, Rumination, Trust Issues **Total: 4**	—
GPT4o-0.7	Blaming, Guilt Tripping, Rumination, Trust Issues **Total: 6**	Minimization, Rumination **Total: 2**

Table 7: List of labels categorized as 'Others' for each model-temperature configuration and prompt type, along with total occurrences.

Cognitive Distortion	GPT4-0.5		GPT4-0.7		GPT4o-0.5		GPT4o-0.7	
	RLP	MLP	RLP	MLP	RLP	MLP	RLP	MLP
All or Nothing Thinking	0.66	0.78	0.60	0.72	0.48	0.62	0.40	0.55
Emotional Reasoning	0.81	0.84	0.76	0.79	0.56	0.49	0.44	0.43
Fortune Telling	0.80	0.834	0.73	0.78	0.59	0.65	0.50	0.58
Labeling	0.87	0.76	0.81	0.72	0.66	0.64	0.56	0.55
Magnification	0.82	0.78	0.75	0.72	0.63	0.66	0.53	0.59
Mental Filter	0.46	0.55	0.42	0.40	0.40	0.26	0.16	0.14
Mind Reading	0.84	0.79	0.80	0.73	0.73	0.68	0.62	0.59
Overgeneralization	0.73	0.74	0.68	0.65	0.58	0.62	0.48	0.51
Personalization	0.81	0.81	0.76	0.77	0.64	0.63	0.54	0.54
Should Statements	0.82	0.74	0.78	0.68	0.75	0.69	0.63	0.59
No Distortion	0.93	0.92	0.91	0.90	0.92	0.89	0.90	0.86
Average	0.78	0.78	0.73	0.71	0.63	0.62	0.52	0.54

Table 8: Fleiss' kappa agreement scores for ten cognitive distortion labels plus the No Distortion label across different GPT-4 models and temperature settings.

the presence of some distortions but not enough agreement to identify it.

Before selecting the final labels, we addressed the presence of additional labels categorized as 'Others' in our dataset. We notice that most of the time in runs, 'Others' was accompanied by at least one CD label from the given list. In such cases, it was simply ignored and further processing was done based on the remaining labels. The 'Others' label was retained only in cases where it was not accompanied by any other CD label.

Table 9 presents the distribution of final cognitive distortion labels obtained from both the Ranked-Label Prompt (RLP) and Multi-Label Prompt (MLP) across four model configurations (GPT-4 and GPT-4o at temperatures 0.5 and 0.7). For each configuration, the table reports both the absolute number of times each label was assigned and the corresponding proportion in percentage in the dataset. As this is a multilabel annotation setting, the to-

tal percentages across labels exceed 100%, since a single input can be assigned multiple distortion labels. The variation in label frequency across configurations highlights the influence of model type, temperature setting, and prompt strategy on annotation behavior.

H. Cognitive Distortion Modeling

Total Model Configurations and Training Experiments: For a comprehensive evaluation, we initially used the transformer-based model MentalRoBERTa, as it has shown strong performance on mental health-related tasks. We also trained MentalBERT for comparison purposes. However, since the performance trends across both models were consistent, we included only MentalRoBERTa results in the main text for clarity. The training details are detailed below:

- **Two models:** MentalBERT and MentalRoBERTa
- **Four configurations:** GPT-4 (T=0.5), GPT-4 (T=0.7), GPT-4o (T=0.5), GPT-4o (T=0.7)
- **Three label sets:** RLP LABELS, MLP LABELS, and GOLDEN LABELS
- **Three classification tasks:**
  - **Binary classification:** In this approach, labels were classified as binary (1 or 0), where 1 indicated the presence of any cognitive distortion, and 0 represented "No Distortion."
  - **Multi-class classification:** Only the dominating RLP LABELS and GOLDEN LABELS were used for this approach, where dominating labels refer to the first label in the case of a multi-label scenario.
  - **Multi-label classification:** The multilabel binarizer was applied to each label column, where each instance could have multiple cognitive distortion labels assigned.

For each configuration considered, the appropriate model was loaded through Hugging Face's Transformer library[7] based model-specific paths, ensuring consistent use of pre-trained parameters. The training was run for maximum of 100 epochs, with early stopping based on F1 score performance. Random initialization variability is taken into account by training for 5 different seed values, with the results averaged across seed values to report mean scores and standard deviations. Overall, the model performance is analyzed across different label sets, classification objectives, and model architectures, providing robust experimental results. Given these factors, the total number of trained models are:

$$\begin{aligned} &2\ \text{(M)} \times 4\ \text{(D)} \times 3\ \text{(L)} \times 1\ \text{(Binary)} \\ &+ 2\ \text{(M)} \times 4\ \text{(D)} \times 2\ \text{(L)} \times 1\ \text{(Multiclass)} \\ &+ 2\ \text{(M)} \times 4\ \text{(D)} \times 3\ \text{(L)} \times 1\ \text{(Multilabel)} \\ &= 24 + 16 + 24 = 64 \text{Models} \end{aligned} \tag{1}$$

Where: M = models, D = configurations based on GPT models and temperatures, L = Label sets.

Since we trained each model using **five different seeds** for robust evaluation, the total number of training runs amounts to:

$$64\ \text{models} \times 5\ \text{seeds} = 320\ \text{trained models} \tag{2}$$

This setup ensures that our findings are statistically robust and account for variability due to random initialization.

I. CD Modeling Results

Table 10 provides a comprehensive breakdown of the F1 (weighted) scores across all model and dataset configurations, classification tasks, and label sets considered in this research. This detailed table complements the main results presented in Table 4 by offering a granular view of performance differences between the MentalBERT and MentalRoBERTa models.

J. Random Baseline ($F1_{random}$)

Theoretical results in Table 11 present the random weighted F1 scores calculated based on the proposed formula in section 5.2 across dataset configurations and classification tasks for the test set(s). We further verify these values through empirical testing.

J.1. Empirical Verification

To verify the accuracy and robustness of our proposed random F1 score calculation methodology, we performed empirical tests by simulation random label assignment using `np.random.choice` method in Python. The verification process involved generating random predictions based on the observed class distributions for each dataset and label type, which were treated as model predictions for calculating the corresponding weighted F1 scores. This process is repeated 1000 times for each configuration, and the averaged weighted F1 values are reported in Table 11 along with the corresponding standard deviation.

[7] https://huggingface.co/

Cognitive Distortion	GPT-4 0.5		GPT-4 0.7		GPT-4o 0.5		GPT-4o 0.7	
	RLP	MLP	RLP	MLP	RLP	MLP	RLP	MLP
All or Nothing Thinking	103 (4.1%)	380 (15.0%)	86 (3.4%)	331 (13.1%)	34 (1.3%)	247 (9.8%)	33 (1.3%)	213 (8.4%)
Emotional Reasoning	959 (37.9%)	947 (37.4%)	932 (36.8%)	886 (35.0%)	296 (11.7%)	119 (4.7%)	202 (8.0%)	110 (4.4%)
Fortune Telling	218 (8.6%)	555 (21.9%)	202 (8.0%)	507 (20.0%)	53 (2.1%)	137 (5.4%)	54 (2.1%)	131 (5.2%)
Labeling	53 (2.1%)	73 (2.9%)	49 (1.9%)	68 (2.7%)	347 (13.7%)	275 (10.8%)	305 (12.1%)	244 (9.6%)
Magnification	274 (10.8%)	298 (11.8%)	234 (9.3%)	252 (10.0%)	481 (19.0%)	570 (22.5%)	437 (17.3%)	532 (21.0%)
Mental Filter	8 (0.3%)	31 (1.2%)	12 (0.5%)	22 (0.9%)	9 (0.4%)	4 (0.2%)	4 (0.2%)	3 (0.1%)
Mind Reading	211 (8.3%)	312 (12.3%)	187 (7.4%)	292 (11.5%)	141 (5.6%)	175 (6.9%)	127 (5.0%)	165 (6.5%)
Overgeneralization	159 (6.3%)	217 (8.6%)	152 (6.0%)	195 (7.7%)	410 (16.2%)	382 (15.1%)	362 (14.3%)	301 (11.9%)
Personalization	656 (25.9%)	713 (28.2%)	597 (23.6%)	676 (26.7%)	250 (9.9%)	338 (13.4%)	201 (7.9%)	301 (11.9%)
Should Statements	45 (1.8%)	80 (3.2%)	39 (1.5%)	74 (2.9%)	28 (1.1%)	61 (2.4%)	24 (1.0%)	54 (2.1%)
Not sure if distortion	64 (2.5%)	81 (3.2%)	99 (3.9%)	105 (4.2%)	83 (3.3%)	138 (5.5%)	113 (4.5%)	183 (7.2%)
Not sure which distortion	28 (1.1%)	1 (0.04%)	30 (1.2%)	1 (0.0%)	143 (5.7%)	88 (3.5%)	284 (11.2%)	165 (6.5%)
Others	2 (0.1%)	0 (0.0%)	2 (0.1%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)
No Distortion	708 (28.0%)	812 (32.1%)	707 (27.9%)	792 (31.3%)	665 (26.3%)	717 (28.3%)	636 (25.1%)	688 (27.2%)

Table 9: Distribution of final labels from both prompts across four model configurations. Values show the number of occurrences of CDs as well as the corresponding proportions.

Labels	Datasets	Binary	Multiclass	Multilabel
MentalRoBERTa				
RLP Labels	Gpt4-0.5	0.838±0.006	0.559±0.011	0.575±0.012
	Gpt4-0.7	**0.854±0.007**	**0.604±0.030**	0.548±0.014
	Gpt4o-0.5	0.832±0.012	0.481±0.017	0.396±0.095
	Gpt4o-0.7	0.809±0.017	0.476±0.020	0.428±0.020
MLP Labels	Gpt4-0.5	0.831±0.010	N/A	**0.609±0.009**
	Gpt4-0.7	0.838±0.009	N/A	0.603±0.010
	Gpt4o-0.5	0.800±0.017	N/A	0.489±0.012
	Gpt4o-0.7	0.829±0.021	N/A	0.474±0.011
Golden Labels	Gpt4-0.5	0.768±0.019	0.384±0.025	0.311±0.018
	Gpt4-0.7	0.770±0.013	0.391±0.021	0.332±0.020
	Gpt4o-0.5	0.778±0.020	0.384±0.016	0.287±0.085
	Gpt4o-0.7	0.813±0.029	0.395±0.028	0.338±0.026
MentalBERT				
RLP Labels	Gpt4-0.5	0.821±0.005	0.533±0.009	0.548±0.017
	Gpt4-0.7	**0.840±0.011**	**0.568±0.014**	0.506±0.018
	Gpt4o-0.5	0.813±0.014	0.441±0.010	0.354±0.028
	Gpt4o-0.7	0.798±0.037	0.454±0.019	0.385±0.019
MLP Labels	Gpt4-0.5	0.809±0.009	N/A	**0.587±0.010**
	Gpt4-0.7	0.828±0.015	N/A	0.570±0.007
	Gpt4o-0.5	0.791±0.009	N/A	0.472±0.003
	Gpt4o-0.7	0.811±0.023	N/A	0.420±0.017
Golden Labels	Gpt4-0.5	0.772±0.014	0.367±0.011	0.320±0.018
	Gpt4-0.7	0.746±0.021	0.369±0.015	0.302±0.021
	Gpt4o-0.5	0.775±0.012	0.359±0.018	0.274±0.076
	Gpt4o-0.7	0.776±0.019	0.365±0.010	0.242±0.049

Table 10: Mean Weighted F1 scores with std on the test sets.

Table 11 shows empirical results closely matched with the theoretical scores, with deviations well within the expected standard error, thus validating our approach. The close alignment between the empirical and theoretical random F1 scores serves as strong evidence that our random F1 score calculation methodology is both accurate and reliable. The verification process also reinforces the validity of using these baseline scores in our kappa calculation (κ_{F1}), enabling a fair comparison of model performance across diverse datasets and label types.

K. Human Verification

Sampling Strategy: We employed a structured sampling strategy for manual verification. This approach aimed to cover the diversity of labels within our dataset while maintaining a fair representation of both frequent and rare labels. We selected annotations generated using the GPT4-0.5 dataset for this process. To construct a well-balanced and meaningful sample for manual verification, we followed a multi-step process:

Label	Configuration	Theoretical Results			Empirical Results (Mean ± Std)		
		Binary	Multiclass	Multilabel	Binary	Multiclass	Multilabel
RLP	Gpt4-0.5	0.589	0.196	0.348	0.588 ± 0.022	0.196 ± 0.019	0.347 ± 0.022
	Gpt4-0.7	0.590	0.204	0.325	0.590 ± 0.023	0.204 ± 0.020	0.325 ± 0.021
	Gpt4o-0.5	0.593	0.178	0.217	0.593 ± 0.022	0.177 ± 0.018	0.217 ± 0.020
	Gpt4o-0.7	0.568	0.184	0.207	0.568 ± 0.022	0.185 ± 0.020	0.208 ± 0.020
MLP	Gpt4-0.5	0.569	N/A	0.455	0.570 ± 0.023	N/A	0.455 ± 0.026
	Gpt4-0.7	0.576	N/A	0.413	0.577 ± 0.023	N/A	0.413 ± 0.024
	Gpt4o-0.5	0.585	N/A	0.253	0.585 ± 0.021	N/A	0.254 ± 0.020
	Gpt4o-0.7	0.568	N/A	0.238	0.568 ± 0.023	N/A	0.238 ± 0.022
Golden	Gpt4-0.5	0.538	0.176	0.201	0.538 ± 0.024	0.176 ± 0.016	0.200 ± 0.018
	Gpt4-0.7	0.539	0.174	0.202	0.537 ± 0.024	0.175 ± 0.017	0.202 ± 0.018
	Gpt4o-0.5	0.547	0.167	0.199	0.547 ± 0.024	0.166 ± 0.018	0.197 ± 0.019
	Gpt4o-0.7	0.548	0.166	0.203	0.548 ± 0.025	0.166 ± 0.017	0.201 ± 0.019

Table 11: Comparison of Theoretical (from 5.2) and Empirical Random F1 Weighted Scores Across Configurations and Classification Tasks for Test Set(s)

1. **Subsetting Based on RLP and MLP Labels:** We created a subset where all labels in the RLP LABELS set were present in the MLP LABELS set, retaining 75.42% of the GPT4-0.5 dataset.
2. **Strict Mismatch Subset Creation:** From the above subset, we identified instances where RLP LABELS and MLP LABELS had no overlap with the *golden labels*, achieving a strict mismatch subset of 35.61% of the original GPT4-0.5 dataset.
3. **Filtering by Maximum Repetition:** To enhance label reliability, we filtered for instances with a maximum label repetition of 5, yielding 31.54% of the original GPT4-0.5 dataset.
4. **Final Sampling Approach:** All instances containing less frequent labels were included entirely, while more frequent labels were capped at 10 instances each. This approach resulted in a final sample size of 101 instances i.e, 3.99% of the original GPT4-0.5 dataset, ensuring a manageable yet meaningful sample set.

Verification Process: We utilized *Label Studio* for conducting the manual verification of our annotated labels. In this process, 3 domain experts were presented with a randomized sample where each instance included:

- **Original Text:** Providing context (USER INPUT) for accurate assessment.
- **Label 1** and **Label 2:** These Labels were shuffled randomly to maintain blinding, ensuring experts could not trace which Label was generated by LLM and which originated from the golden standard.

Experts were asked to select which label they agree with using the following options: a) **Label 1**, b) **Label 2**, c) **Both Labels**, d) **None of the Labels**, e) **Partial Label 1** (if applicable), and f) **Partial Label 2** (if applicable). This blinded evaluation method promotes unbiased feedback, allowing us to assess the quality of our model's annotations against expert judgment.