Signs of Struggle

Spotting Distorted Thoughts in Social Media Text

by

Abhishek Kuber

to obtain the degree of Master of Science in Computer Science
at the Delft University of Technology,
to be defended publicly on Monday, 23 June 2025 at 13:00.

Student Number:	5966531
Project Duration:	November, 2024 - June, 2025
Faculty:	Electrical Engineering, Mathematics and Computer Science
Department:	Interactive Intelligence
Thesis Advisor:	Dr. P.K. Murukannaiah (EEMCS)
Daily Supervisors:	Dr. Ir. E. Liscio (EEMCS)
	Ir. R. Zhang (TPM)
Thesis Committee:	Dr. J. Yang (EEMCS)

TU Delft

Acknowledgements

In a way, this is probably the most meaningful part of my thesis, and the hardest to write. I'm deeply grateful to the people who have supported me throughout this journey. Without their guidance, encouragement, and presence, I wouldn't have made it this far.

First and foremost, I would like to thank my supervisor, Dr. Enrico Liscio. Your detailed feedback and thoughtful questions shaped every part of this project, and the document you are reading right now wouldn't be what it is if not for his feedback. Even when I felt stuck, you helped me find a path forward. Thank you for teaching me how to think critically and work rigorously. Grazie mille!

I'm also thankful to Dr. Pradeep Murukannaiah and Ruixuan Zhang for their valuable insights. Your perspectives helped me approach this project from different angles. Thank you, Pradeep, for your guidance on the broader direction of the thesis and for helping me see how it could grow. Thank you, Rae, for offering a social science lens, for reviewing my work, and for your help with the annotations. In the early stages of the project, Enrico, Pradeep, and Ruixuan helped me refine the scope and shape it into something manageable. I feel fortunate to have had such knowledgeable mentors steering me in the right direction.

To my parents, thank you for supporting me unconditionally and encouraging me to follow my dreams, even if it meant me moving thousands of miles away. Aai, Baba, Aditi, Ajji and Coco - thank you for always being there. Your love, support, and calls from home meant the world to me. Your belief in me gave me the strength to pursue this path.

To Shiwangi - thank you for being my sounding board through every rant, for helping me rehearse presentations, for reading my drafts, and most of all, for your patience. Thank you to my desk partner Paul, for the lunches, the siestas in the sun, and the chats. Even though we were working on different projects, it felt like we were in it together. And finally, to Zoë and Saga - you made life in Delft memorable. Our weekends and little trips gave me the breathers I didn't know I needed. I truly couldn't have survived Delft without you.

Thank you, again and again, to everyone who made this journey possible.

Abhishek Kuber
Delft, June 2025

Signs of Struggle: Spotting Distorted Thoughts in Social Media Text

Abhishek Kuber
Delft University of Technology, The Netherlands

Abstract

Rising mental health issues among adolescents have increased interest in automated approaches for detecting early signs of psychological distress in digital text. One important focus is the identification of cognitive distortions – irrational thought patterns – because of their role in aggravating mental distress, and early detection may enable timely, low cost interventions. While prior work has focused on English data, we present a first in-depth study of cross lingual and cross register generalization for cognitive distortion detection, using forum posts written by Dutch adolescents. We frame the task at two levels: (1) detecting whether a post contains a cognitive distortion, and (2) identifying the specific text span that expresses it. Our findings show that domain adaptation methods perform best for post-level detection, while a simpler technique – sentence embeddings with a classifier – outperforms more complex models for span identification. Results show predicting cognitive distortions in text is challenging, and highlight how changes in language and writing style can significantly impact performance.

1 Introduction

Mental health disorders among adolescents are a growing global concern. According to the World Health Organization, one in seven individuals aged 10-19 experiences a mental disorder, with depression, anxiety and behavioural disorders being the most common ones (World Health Organization, 2024). This trend is mirrored in The Netherlands, with approximately 13% of girls and 7% of boys aged 12 to 17 reporting experiencing mental health problems in 2023 (Statistics Netherlands, 2024).

Despite rising awareness, many cases go undetected and untreated. Adolescence is a critical period of emotional and cognitive development, where unaddressed conditions can have lasting effects into adulthood, highlighting the need for early, non-pharmacological interventions (World Health Organization, 2024). A common method in digital mental health research is analyzing social media data, which captures authentic expressions of emotion and help-seeking behavior (Chancellor and De Choudhury, 2020).

With 96% of the adolescent population using the internet daily (ACT for Youth, 2024), digital platforms have become key spaces where mental health struggles are voiced. One such platform is De Kindertelefoon[1], where Dutch youth aged 8-18 can discuss on issues such as sexuality, bullying and emotional struggles on anonymous forums. As a rich source of self reported experiences, the forum offers valuable insights into youth mental health and emotional distress, and provides a unique opportunity to explore automated techniques for understanding and supporting adolescent mental well-being.

The foundation for these techniques comes from Cognitive Behavioral Therapy (CBT), developed by Aaron Beck (Beck, 1970). CBT is a widely used treatment for mental health disorders (Curtiss et al., 2021; David et al., 2018b), and it emphasizes that our interpretations of events – not the events themselves – determine how we feel. For instance, viewing a breakup as "No one will ever love me again", over time, may lead to social withdrawal and loneliness. Reinforcing these negative thought patterns, known as *cognitive distortions*, can aggravate mental distress and are linked to conditions like depression and anxiety (Beck, 1970; Persons et al., 2023). A key CBT technique to combat this is cognitive reframing (see Appendix A.2), which begins by identifying distorted thoughts, making automated detection a necessary first step.

Prior work has primarily focused on English data. In this study, we perform the first in-depth study of how computational methods for cognitive distortion detection generalize across both language

[1] https://forum.kindertelefoon.nl/

and register. Here, register refers to variation in language across groups, in our case, a shift from adult written English data to adolescent written Dutch texts, which presents new linguistic challenges. We frame the task at two levels: (1) detecting whether a post contains a distortion, and (2) locating the specific span expressing it. While the first is more straightforward, the second supports better cognitive reframing. Our experiments range from prompting to supervised learning and domain adaptation, evaluating generalization across both language and register shifts.

Our results show that while multilingual models can generalize across languages, they often struggle with register changes such as writing style. Domain adaptation proves essential for improving performance at the post level. At the sentence level, we find that simpler sentence embedding methods outperform more complex approaches. Overall, our work demonstrates that cognitive distortion detection can be adapted across languages and registers – a critical first step towards making them more generalizable.

2 Background and Related Work

2.1 Computational Approaches to CBT

Early approaches to detecting cognitive distortions in text relied on linguistic features and classical classifiers. For instance, Simms et al. (2017) used LIWC features from Tumblr posts for binary classification, while Shreevastava and Foltz (2021) compare semantic and syntactic feature types, and show that combining Sentence-BERT embeddings with SVMs improves performance.

With the rise in popularity of transformer models, supervised learning has gained traction. Jiang et al. (2024) framed distortion detection as a hierarchical classification task using ERNIE 3.0 – a supervised transformer model pretrained on knowledge graphs – to label ABCD components. They then applied Pegasus and GPT-4 for summarization and reasoning, but reported hallucination issues with LLMs, highlighting the reliability of supervised methods.

Prompt based approaches have also become popular. For example, Chen et al. (2023) introduced Diagnosis of Thought prompting, grounded in cognitive theory, though it tended to overpredict distortions. Lim et al. (2024) proposed ERD, combining extraction and debate across multiple LLMs, improving precision via a judge model. Chatbot based systems like TeaBot (Nazarova, 2023) use GPT-3 for real-time distortion detection and challenging distortions using CBT inspired questions.

Recent work expands beyond detection to include reframing. Sharma et al. (2023) used few-shot prompting with GPT-3 to generate reframed thoughts, guided by linguistic cues. ReframeGPT (Wang et al., 2024) and RESORT (Zhan et al., 2024) are frameworks that iteratively refine reframes along different linguisitic dimensions and coping strategies.

Building on these strategies, we combine contextual embeddings with lexical features to improve classification performance. We also evaluate a range of methods that utilize LLMs – prompting, instruction tuning, and supervised finetuning for both classification and span detection. However, as shown in Section 4.1, they do not generalize across registers, hence, highlighting the need for domain adaptation techniques.

2.2 Domain Adaptation

In NLP, a domain typically refers to a coherent corpus shaped by topic, style, or language use. Domain adaptation tackles the challenge of applying models trained on one domain to another, often facing performance drops due to such variations (Plank and Van Noord, 2011; Ramponi and Plank, 2020). To overcome this, various strategies have been proposed to improve cross domain generalization.

For example, contrastive learning mitigates this by pulling semantically similar examples (e.g., augmentations or same-label pairs) closer and pushing dissimilar ones apart, improving feature robustness across domains (Li et al., 2020; Luo et al., 2022; Gao et al., 2022). Adversarial training aims to learn domain invariant features by confusing the model's ability to identify the input domain (Du et al., 2020; Ganin et al., 2016; Liu et al., 2017; Zhou et al., 2020).

Long et al. (2022) propose Domain Confused Contrastive Learning (DCCL), which encourages the model to discard domain specific cues via domain puzzles. We build on this idea, using it to jointly tackle differences across language and register, aiming to learn representations that generalize across both language and writing style.

We adopt DCCL due to its suitability for our scenario involving register shifts. Unlike other adversarial methods, DCCL introduces intermediate "domains" through the domain puzzles, enabling smoother transitions between source and target dis-

tributions. Moreover, by masking domain-specific cues through the puzzles, DCCL forces the encoder to focus on domain invariant, yet task relevant semantics.

3 Datasets

3.1 De Kindertelefoon (KT)

De Kindertelefoon is a Dutch organization that has supported children and adolescents since 1979, initially as a helpline offering a safe, anonymous space to discuss personal problems. Over time, it has evolved into a broader platform offering various forms of support, including online chat services and moderated forums. These forums allow young people aged 8-18 to express their thoughts, ask questions and seek advice on topics such as bullying, sexuality, relationships and mental health – enabling a unique form of peer-to-peer support.

We collect 37,691 public posts across the 16 subforums. Data is anonymized and usernames are pseudonymized to remove personally identifiable information (see Appendix A.1). This was approved by the Human Research Ethics Committee of the Delft University of Technology (project number 5545).

3.1.1 Annotation Process

Two annotators were asked to manually label 100 randomly selected posts from the *emotionele problemen en gevoelens* (emotional problems and feelings) subforum. They were tasked with two tasks: (1) assign a binary label indicating whether the post contains a cognitive distortion, and (2) if a distortion is present, identify the specific sentence(s) expressing the distorted thinking.

Before starting the task, annotators were provided with detailed annotation guidelines (see Appendix A.3). These include a definition of cognitive distortions as irrational or negative, biased thought patterns that distort a person's perception of reality and can contribute to emotional distress. Ten common distortion types were provided as reference, each with a brief description and an example. Annotators were instructed to only label a post as "Yes" if the content clearly matched one of the defined distortion types, to avoid over interpretation, and to rely solely on information explicitly stated in the text. They were also asked to mark full sentences rather than spans.

After completing the task independently, inter-annotator agreement was computed using Cohen's Kappa, yielding $\kappa = 0.52$, indicating moderate agreement. The annotators then discussed disagreements and resolved them through deliberation. Upon reaching consensus, it resulted in an improved agreement of $\kappa = 0.88$.

Following this process, the author of this paper annotated an additional 350 posts with binary labels (Yes/No) at the post level to expand the dataset for training and evaluation.

3.2 Therapist Q&A (EN)

Shreevastava and Foltz (2021) release an annotated dataset based on user-submitted mental health queries in English, each originally answered by licensed therapists. The dataset labels each entry as either containing a specific cognitive distortion or no distortion, and includes the corresponding span of distorted text. Since, to the best of our knowledge, no comparable annotated dataset exists in Dutch, we incorporate this English dataset into model training, evaluating the generalizability on different test sets.

3.3 Translated Therapist Q&A (NL)

We translate the English dataset into Dutch to isolate the effect of language change. Since both the EN and NL datasets share a similar register, this allows us to examine how well models trained on English data generalize across language alone – without the influence of variations in register.

4 Stage 1 : Distortion Detection

4.1 Establishing a baseline

To establish a baseline for cross lingual knowledge transfer, we first perform a series of experiments using only the EN data for training. All models are evaluated on three test sets – EN, NL and KT. While evaluating on KT data involves both a change in language and register, testing on NL data isolates the impact of language alone by keeping the register constant. We try out the following methods:

- **Adapters:** Adapters (Houlsby et al., 2019) are lightweight, plug-in modules inserted into pretrained transformer layers that enable efficient finetuning for new tasks without updating the full model. We use XLM-RoBERTa with adapters for binary sequence classification via AdapterHub (Pfeiffer et al., 2020).
- **XLMR Finetuning:** We finetune XLM-RoBERTa with a binary classification head

on top.

- **Prompting LLaMA:** Prompting is tested in two ways - using a short, instruction only prompt (Appendix A.4.1), and a long prompt which includes definitions and examples (Appendix A.4.2).
- **LLaMA Instruction Tuning:** We finetune LLaMA using instruction tuning with the short prompt (Appendix A.4.1).
- **LLaMA Finetuning:** We finetune LLaMA with a binary classification head on top.

The 5 fold cross validated results are presented in Table 1. Since the goal is to test cross lingual generalization, we only report the weighted F1 scores for each dataset.

Method	EN	NL	KT
Adapters	0.74 ± 0.02	**0.73 ± 0.01**	**0.56 ± 0.04**
XLMR Finetuning	0.74 ± 0.01	**0.73 ± 0.03**	0.54 ± 0.08
LLaMA (SP)	0.61 ± 0.02	0.62 ± 0.02	0.39 ± 0.06
LLaMA (LP)	0.59 ± 0.02	0.61 ± 0.03	0.46 ± 0.04
LLaMA Finetuning	**0.77 ± 0.08**	0.71 ± 0.08	0.51 ± 0.08
LLaMA I.T.	0.63 ± 0.03	0.61 ± 0.05	0.50 ± 0.06
Random Baseline	0.50 ± 0.02	0.51 ± 0.00	0.52 ± 0.05

Table 1: Weighted F1 scores for preliminary distortion detection methods. All methods have been trained on English data (EN), and tested on English data (EN), English data translated to Dutch (NL), and De Kindertelefoon posts (KT). For LLaMA based methods, SP=Short Prompt, LP=Long Prompt, I.T.=Instruction Tuning.

From Table 1, we see that LLaMA Finetuning achieves the highest F1 on the EN set (0.77), followed by Adapters and XLM-RoBERTa Finetuning (0.74). Similarly, on the NL set, XLM-RoBERTa Finetuning leads along with Adapters (0.73), with LLaMA Finetuning (0.71) close behind. Performance drops notably on the KT set, with Adapters performing the best (0.56), while the other models fall between 0.39 and 0.54.

These results suggest that knowledge transfer across languages is not the problem, but rather it is the differences in register. EN data is comprised of texts from adults seeking guidance on an online therapy platform, explaining their problems in detail. In contrast, KT texts feature Dutch adolescents sharing their thoughts informally with peers, often with less elaboration and the use of slang. These differences in register likely affect the models' ability to generalize, highlighting the need for techniques beyond simple cross lingual transfer learning.

Given the comparable performance between LLaMA and XLM-RoBERTa classifiers, we choose XLM-RoBERTa for subsequent experiments. As an encoder-only model, XLM-RoBERTa is better suited for classification tasks as compared to the decoder-only LLaMA. Additionally, given its smaller size, XLM-RoBERTa is more sample efficient, requiring less data and compute for finetuning.

4.2 Improving Generalization

Building on the findings from our preliminary baseline experiments, we explore a set of approaches aimed at improving generalization to the KT data.

4.2.1 Rewriting

Our first approach involves rewriting texts from the English dataset. Specifically, we prompt *meta-llama/Llama-3.1-8B-Instruct* to rewrite sentences from the English dataset as a Dutch teenager on De Kindertelefoon (see Appendix A.4.3). We then use this dataset to finetune XLM-RoBERTa.

4.2.2 Incorporating Lexical Features

Prior work shows that lexical features add useful information to contextual embeddings by capturing signals like emotion and topic that pretrained models may miss. Based on this, we combine lexical features with embeddings for classification.

We use Empath (Fast et al., 2016) to extract 195 lexical features from KT posts. A paired t-test identifies 68 features that differ significantly between distorted and non-distorted texts (see Appendix A.4.4). For classification, we encode the text through sentence embeddings, apply mean pooling to the last hidden state to get a representation, concatenate the selected Empath features to it, and feed the resulting embedding into a classifier.

4.2.3 Domain Confused Contrastive Learning

Introduced by Long et al. (2022), this method encourages the model to learn domain invariant, task discriminative representations by introducing domain puzzles. These puzzles take the form of learnable perturbations added to the embeddings, and promote generalization across domains. The architecture of the model is shown in Figure 1.

We add these perturbations to post embeddings and pass them through a domain classifier trained to distinguish between EN (source) and KT (target) texts. These perturbations are optimized to confuse

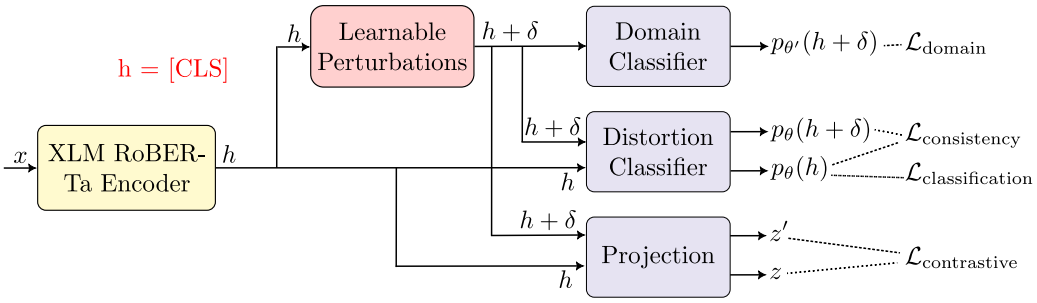


Figure 1: Architecture for Domain Confused Contrastive Learning

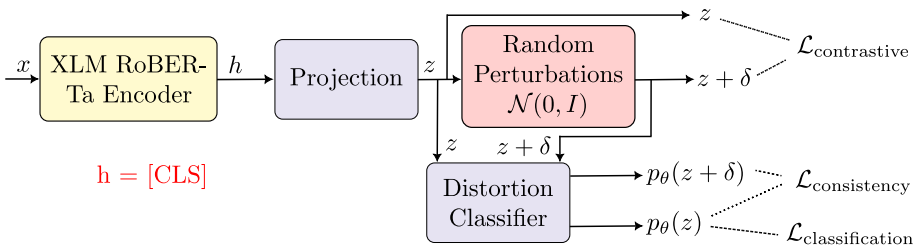


Figure 2: Architecture for Naive Domain Confused Contrastive Learning

the classifier, encouraging the model to discard domain specific cues and generalize better. The resulting embeddings lie in an intermediate domain – a shared space between the source and target domains – enabling better alignment. To achieve this, the domain classification loss ($\mathcal{L}_{\text{domain}}$) is maximized.

Both the original and perturbed embeddings are passed through a down-projection layer, which helps remove redundant information and highlights meaningful features. To bring the projected embeddings closer in the embedding space, a contrastive loss (InfoNCE[2]) is applied on the original and perturbed projected embeddings ($\mathcal{L}_{\text{contrastive}}$).

The distortion classifier processes both the original and perturbed embeddings. However, only the original embeddings are used for predicting whether the text contains a distortion or not ($\mathcal{L}_{\text{classification}}$). To ensure that the model's predictions remain consistent despite the perturbations, we impose a consistency loss (Kullback-Leibler divergence) between the logits of the original and perturbed embeddings ($\mathcal{L}_{\text{consistency}}$). The final loss is given by:

$$\mathcal{L} = \alpha \cdot \mathcal{L}_{\text{domain}} + \beta \cdot \mathcal{L}_{\text{consistency}} +$$

$$\lambda \cdot \mathcal{L}_{\text{contrastive}} + \mathcal{L}_{\text{classification}},$$

where $\alpha = 1e-3, \beta = 5, \lambda = 3e-2$ are the coefficients for the losses, taken from Long et al. (2022). There are two training loops for this model: the first incorporates all losses, while the second uses only the classification loss, updating only its associated components.

[2] https://github.com/RElbers/info-nce-pytorch

4.2.4 Naive Domain Confused Contrastive Learning

To examine whether contrastive learning alone is sufficient to encourage domain invariance, we experiment with a simplified variant of DCCL that removes the domain classifier and learnable perturbations. Each loss serves the same purpose as in DCCL; however, instead of learning perturbations, this method applies *random Gaussian noise*, enforcing domain invariance *without needing an adversarial domain classifier*. Despite the architectural differences, the objective remains the same, trying to learn domain invariant representations. The training method is same as DCCL, with two training loops. The architecture for this model can be seen in Figure 2.

4.3 Results

The models are trained on different data configurations, and the results are shown in Table 2. The weighted Precision, Recall, and F1 scores are reported for each method.

Table 2 shows that prompting based methods (LLaMA SP and LP) perform poorly, while even limited supervision on KT data improves results. DCCL achieves the highest score, followed closely by Naive DCCL, Adapters, and

Data	Method	Precision	Recall	F1
None	LLaMA SP	0.55 ± 0.07	0.45 ± 0.05	0.39 ± 0.06
	LLaMA LP	0.56 ± 0.05	0.48 ± 0.04	0.46 ± 0.04
	Random Baseline	0.50 ± 0.06	0.48 ± 0.06	0.48 ± 0.06
EN	Adapters	**0.76 ± 0.02**	0.59 ± 0.03	0.56 ± 0.04
	LLaMA Finetuning	0.53 ± 0.11	0.57 ± 0.09	0.51 ± 0.08
	XLMR Finetuning	0.73 ± 0.06	0.57 ± 0.06	0.54 ± 0.08
	LLaMA I.T.	0.57 ± 0.10	0.50 ± 0.07	0.50 ± 0.06
RE-EN	XLMR Finetuning	0.73 ± 0.05	0.54 ± 0.06	0.49 ± 0.10
EN + KT	Adapters	0.67 ± 0.03	0.67 ± 0.03	0.67 ± 0.05
	LLaMA Finetuning	0.61 ± 0.04	0.61 ± 0.04	0.58 ± 0.04
	XLMR Finetuning	0.47 ± 0.25	0.58 ± 0.11	0.46 ± 0.16
	LLaMA I.T.	0.64 ± 0.04	0.64 ± 0.04	0.64 ± 0.04
	XLMR+Empath	0.70 ± 0.06	0.69 ± 0.06	0.69 ± 0.07
	Naive DCCL	**0.74 ± 0.03**	0.69 ± 0.08	0.67 ± 0.14
	DCCL	0.74 ± 0.05	**0.73 ± 0.04**	**0.73 ± 0.05**

Table 2: Weighted Precision, Recall, and F1 Scores for Stage 1 methods. "Data" indicates the training set: None, EN (original English), RE-EN (rewritten in De Kindertelefoon style), and KT (De Kindertelefoon). For LLaMA based methods, SP=Short Prompt, LP=Long Prompt, I.T.=Instruction Tuning. All models are tested on KT. Best results are bold based on McNemar's test ($p < 0.05$); statistically insignificant results are underlined (see Appendix A.4.5).

XLMR+Empath. Notably, the performance gain from XLMR+Empath over standard finetuning highlights the added value of lexical features in distinguishing distorted from non-distorted text.

To understand why DCCL performs better than the other methods, we visualize the embedding space of DCCL and Naive DCCL using UMAP (McInnes et al., 2020), projecting the embeddings to 2D. Figure 3 shows the resulting projections of the embeddings for the EN and KT texts.

Figure 3 shows how the methods structure the embedding space. For XLM-RoBERTa (Figures 3a, 3b), distorted and non-distorted texts overlap heavily in the EN and KT embedding spaces, showing minimal separation. DCCL (Figures 3g, 3h) achieves clearer separation, indicating it captures distortion relevant features. Naive DCCL (Figures 3d, 3e) attempts the separation but it remains less pronounced.

When examining the embeddings of both EN and KT posts plotted together (Column 3), XLM-RoBERTa (Figure 3c) exhibits an obvious language divide – EN and KT posts are clustered in clearly separated regions. In contrast, both DCCL and Naive DCCL reduce this separation, suggesting an effort to learn distortion specific, domain invariant features. However, Naive DCCL's embeddings (Figure 3f) remain less structured than DCCL's (Figure 3i).

Table 2 supports these findings. While Naive DCCL improves over XLM-RoBERTa, only DCCL – with learnable perturbations and domain confusion – achieves the best performance, contrastive learning alone appears insufficient.

5 Stage 2 : Identifying the distorted span

The goal of this stage is to identify the specific distorted segment within a post, which we cast as a sentence classification task. As a starting point, we prompt LLaMA to directly identify distorted spans, without any finetuning. We then investigate whether the improved domain alignment observed in Stage 1 carries over to this finer grained task. Specifically, since both DCCL and Naive DCCL brought EN and KT post embeddings closer in the representation space, we evaluate whether their encoders can support accurate span identification.

5.1 Methods

5.1.1 Prompting LLaMA

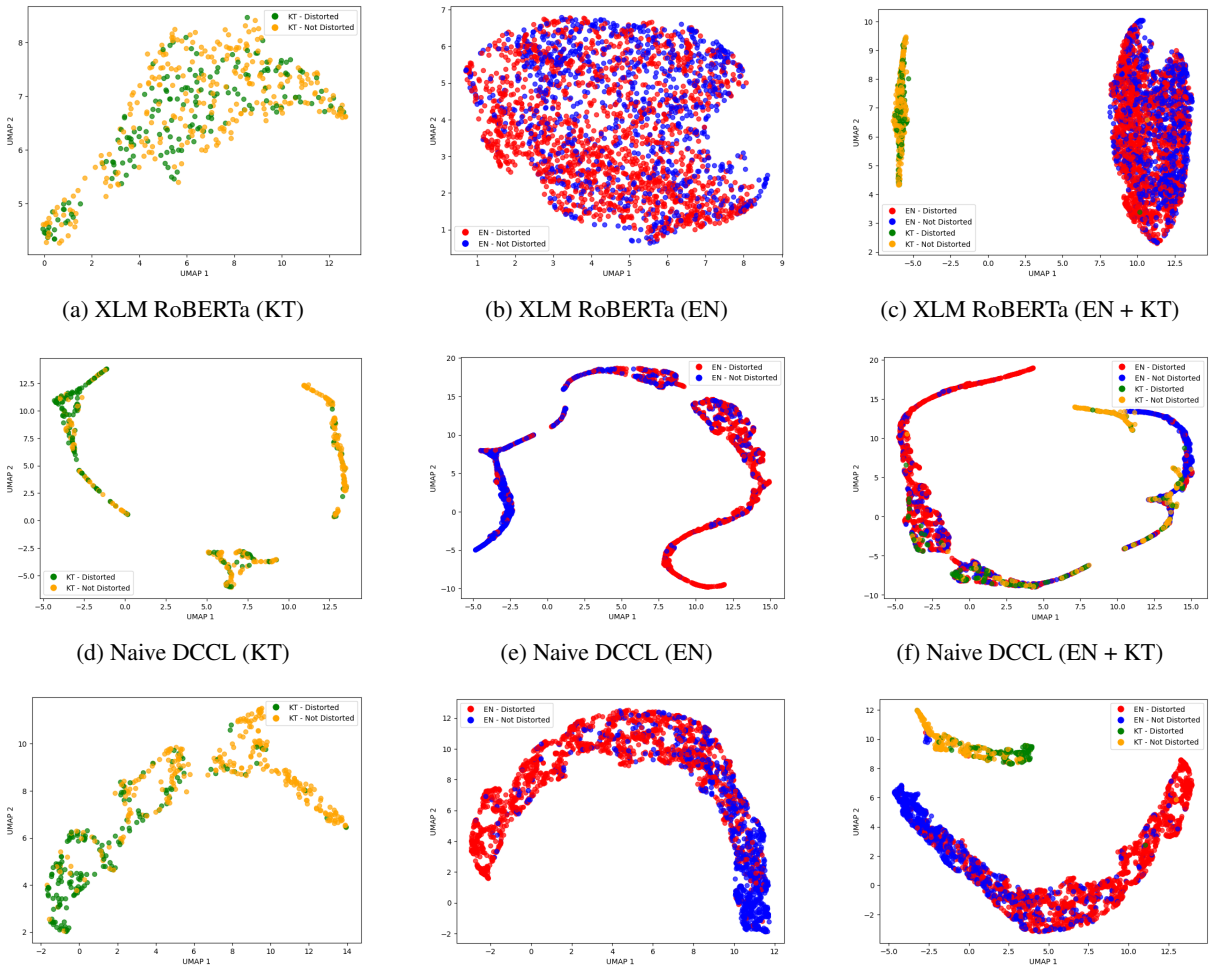
We begin by prompting LLaMA directly, exploring two configurations – unstructured output and structured output.

In the unstructured output setting, the model is given the full post as a paragraph and asked to return the sentences it considers distorted, also formatted as a paragraph (see Appendix A.5.1). These predicted sentences are then matched against the original post to produce a binary sequence.

In the structured output approach, the post is split into individual sentences before being passed to the model. The model is prompted to output the corresponding binary sequence directly (see Appendix A.5.2).

5.1.2 Encoder + Classifier

Since this is a sentence classification task, we pass each sentence independently through an encoder



(a) XLM RoBERTa (KT) (b) XLM RoBERTa (EN) (c) XLM RoBERTa (EN + KT)

(d) Naive DCCL (KT) (e) Naive DCCL (EN) (f) Naive DCCL (EN + KT)

(g) DCCL (KT) (h) DCCL (EN) (i) DCCL (EN + KT)

Figure 3: UMAP plots of the embeddings for XLM RoBERTa, DCCL and Naive DCCL. Column 1 represents embeddings of De Kindertelefoon (KT) texts, Column 2 corresponds to embeddings of English (EN) texts, and Column 3 shows both combined. Row 1 displays embeddings from XLM RoBERTa, Row 2 from Naive DCCL, and Row 3 from DCCL. The yellow dots represent non distorted KT posts, the green represent distorted KT posts, the blue represent non distorted EN texts and red represent distorted EN texts.

to obtain a representation, which is then used to classify the sentence as distorted or not distorted.

We experiment with three encoder variants: XLM-RoBERTa, DCCL, and Naive DCCL. For embedding strategies, we evaluate two options: the [CLS] token and mean pooling over the last hidden layer. Each embedding is then passed to either a linear classifier or a SVM for the final prediction.

In the linear classifier setup, both the encoder and classifier are trained jointly in an end-to-end fashion. In contrast, since SVMs are non differentiable, the encoder is frozen and used only to get embeddings, which are then classified by the SVM.

5.2 Results

Since the data in this stage has a large class imbalance (see Appendix A.1.1), we report the weighted F1 score to provide a single overall measure while also presenting the per class F1 scores. This allows for a more detailed assessment of how well the model performs on both majority and minority classes. Results are presented in Table 3.

Looking at the results from Table 3, the highest performance is achieved by a simple model: an SVM classifier with a RBF kernel applied to mean pooled embeddings from a XLM-RoBERTa encoder, yielding a weighted F1 score of 0.75. Prompting based methods with LLaMA perform poorly, achieving relatively high weighted F1 scores around 0.84 due to class imbalance, but failing to correctly identify distorted texts. Similarly, both LLaMA and encoders with linear classifiers tend to overpredict the non-distorted class.

Interestingly, while DCCL was the best perform-

Method	Class 0	Class 1	Weighted
LLaMA (U)	0.91 ± 0.02	0.27 ± 0.10	0.84 ± 0.04
LLaMA (S)	0.92 ± 0.03	0.25 ± 0.07	0.84 ± 0.05
BCL	0.95 ± 0.01	0.17 ± 0.10	**0.90 ± 0.00**
BML	0.96 ± 0.00	0.07 ± 0.07	0.89 ± 0.00
NCL	0.96 ± 0.00	0.19 ± 0.13	0.90 ± 0.01
NML	0.96 ± 0.00	0.08 ± 0.10	0.90 ± 0.01
DCL	0.96 ± 0.00	0.14 ± 0.11	0.90 ± 0.01
DML	0.96 ± 0.00	0.08 ± 0.12	0.90 ± 0.01
BCS	0.70 ± 0.02	0.72 ± 0.03	0.71 ± 0.02
BMS	0.74 ± 0.07	**0.76 ± 0.07**	0.75 ± 0.07
NCS	0.64 ± 0.07	0.63 ± 0.06	0.64 ± 0.06
NMS	0.66 ± 0.05	0.67 ± 0.06	0.67 ± 0.05
DCS	0.67 ± 0.04	0.66 ± 0.06	0.66 ± 0.05
DMS	0.70 ± 0.05	0.72 ± 0.04	0.71 ± 0.04

Table 3: F1 scores for each model configuration, reported separately for Class 0, Class 1, and the weighted average. Each method is represented by three letters, except for the two LLaMA based methods: LLaMA (U) means prompting with the unstructured prompt, and LLaMA (S) means the structured prompt. For the remaining methods, the first letter indicates the encoder (B = XLM-RoBERTa, D = DCCL, N = Naive DCCL), the second letter the embedding type (C = [CLS] token, M = mean pooled), and the third letter the classifier (L = linear classifier, S = SVM). For SVM models, only the best-performing kernel is shown; all SVM results are provided in Table 12.

ing model in Stage 1, it does not perform as well in Stage 2. We conjecture that this drop in performance could be due to the nature of the task. DCCL is designed to work at the post level, where there is more domain specific information visible – such as writing style, slang, or structure – that helps the model learn useful perturbations. But at the sentence level, these patterns are much less obvious. As a result, DCCL and Naive DCCL might not add much value and could even introduce unnecessary noise. Additionally, since the same perturbations are used without retraining for sentence level inputs, it could explain why their performance drops.

Finally, across most configurations, mean pooling consistently outperforms the [CLS] token. This suggests that aggregating information across all tokens provides a more stable and representative signal for identifying distorted language. Since distorted language can be subtle and context-dependent, relying on one token to capture these cues might lead to bad results. Mean pooling, by capturing signals from all parts of the input, offers a better embedding that appears better suited to this task.

Overall, these findings highlight that improvements in domain alignment at one level do not always extend to finer levels of granularity, and that model simplicity can often outperform complexity when tasks change in nature or scale.

6 Conclusion

This study explores the automatic detection of cognitive distortions in Dutch adolescent social media posts, a critical first step toward AI-assisted cognitive reframing. To the best of our knowledge, this is the first study to evaluate both cross lingual and cross register generalization. We develop a two-stage framework that: (1) detects whether a post contains a cognitive distortion, and (2) identifies the exact distorted span within the post.

We compare a range of approaches, from LLMs to simpler classifiers. At the post level, we demonstrate that domain adaptation is essential for generalization across registers, with our approach aligning representations between English and De Kindertelefoon data. At the sentence level, however, a simpler setup – mean-pooled embeddings from a XLM-RoBERTa encoder paired with a SVM classifier – yields the best results. Across both tasks, prompt based methods yield notably lower performance, reinforcing the conclusion from Jiang et al. (2024) that supervised methods remain more effective in this context. Our findings also highlight how domain shifts across age, language, and writing style can significantly impact performance.

Future work should focus on identifying the "Activating Event" and "Consequence" components to complete the ABC model and enable effective cognitive reframing. Due to time constraints, we did not explore domain adaptation for Stage 2 – aligning sentence level representations across registers remains an open challenge.

Limitations

While our results are promising, there remain several avenues for improvement. First, De Kindertelefoon dataset is only partially annotated. Although inter annotator agreement improves significantly after deliberation, the limited volume of labeled data may be constraining model performance. Expanding the annotation effort through techniques

such as active learning, or simply getting more annotators, could potentially boost performance, as results show that training on a few examples from De Kindertelefoon dataset gives better performance. Moreover, incorporating annotations verified by mental health professionals may further enhance reliability and clinical validity.

Currently, distorted spans are identified at the sentence level. While this approach is simple and practical, it may overlook subtler cues embedded within sentences. Moving toward finer grained methods, such as token level span prediction or sequence tagging, could significantly enhance both precision and interpretability.

Another technical limitation involves handling long posts. At present, inputs longer than 512 tokens are truncated as that is the maximum context length of XLM-RoBERTa, potentially omitting important context. Exploring multilingual models with longer context lengths may help capture dependencies in forum posts more effectively, potentially improving performance.

Ethical Considerations

The use of AI for detecting and reframing cognitive distortions in children's text raises important ethical questions. First, since the nature of the data is sensitive, there must be data protection laws in place to prevent misuse or accidental disclosure. Second, while AI can offer helpful cognitive reframing suggestions, adolescents may become frustrated or distressed by repetitive interventions, highlighting the need for carefully designed user experiences. Third, it should never replace trained professionals, rather, it must be thought of as a tool that supports trained mental health professionals.

Acknowledgments

Research reported in this work was partially or completely facilitated by computational resources and support of the Delft AI Cluster (DAIC) at TU Delft (RRID:SCR_025091), but remains the sole responsibility of the authors, not the DAIC team.

References

ACT for Youth. 2024. Accessed: 01-05-2025. [link].

Krishna C. Bathina, Marijn ten Thij, Lorenzo Lorenzo-Luaces, Lauren A. Rutter, and Johan Bollen. 2020. Depressed individuals express more distorted thinking on social media. *Preprint*, arXiv:2002.02800.

Aaron T. Beck. 1970. Cognitive therapy: Nature and relation to behavior therapy. *Behavior Therapy*, 1(2):184–200.

Stevie Chancellor and Munmun De Choudhury. 2020. Methods in predictive techniques for mental health status on social media: a critical review. *NPJ Digital Medicine*, 3(1):43.

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

Gavin Clark and Sarah Egan. 2015. The socratic method in cognitive behavioural therapy: A narrative review. *Cognitive Therapy and Research*, pages 1–17.

Joshua E. Curtiss, Daniella S. Levine, Ilana Ander, and Amanda W. Baker. 2021. Cognitive-behavioral treatments for anxiety and stress-related disorders. *FOCUS*, 19(2):184–189.

Michael Han Daniel Han and Unsloth team. 2023. Unsloth.

Daniel David, Carmen Cotet, Silviu Matu, Cristina Mogoase, and Simona Stefan. 2018a. 50 years of rational-emotive and cognitive-behavioral therapy: A systematic review and meta-analysis. *Journal of Clinical Psychology*, 74(3):304–318.

Daniel David, Ioana Cristea, and Stefan G. Hofmann. 2018b. Why cognitive behavioral therapy is the current gold standard of psychotherapy. *Frontiers in Psychiatry*, 9.

Chunning Du, Haifeng Sun, Jingyu Wang, Qi Qi, and Jianxin Liao. 2020. Adversarial and domain-aware BERT for cross-domain sentiment analysis. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4019–4028, Online. Association for Computational Linguistics.

Olive Jean Dunn. 1961. Multiple comparisons among means. *Journal of the American Statistical Association*, 56(293):52–64.

Albert Ellis. 1957. Rational psychotherapy and individual psychology. *Journal of individual psychology*, 13(1):38.

Albert Ellis. 1991. The revised abc's of rational-emotive therapy (ret). *Journal of Rational-Emotive and Cognitive-Behavior Therapy*, 9(3):139–172.

Ethan Fast, Binbin Chen, and Michael S. Bernstein. 2016. Empath: Understanding topic signals in large-scale text. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, CHI'16, page 4647–4657. ACM.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario March, and Victor Lempitsky. 2016. Domain-adversarial training of neural networks. *Journal of Machine Learning Research*, 17(59):1–35.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2022. Simcse: Simple contrastive learning of sentence embeddings. *Preprint*, arXiv:2104.08821.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin de Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for nlp. *Preprint*, arXiv:1902.00751.

Meng Jiang, Yi Jing Yu, Qing Zhao, Jianqiang Li, Changwei Song, Hongzhi Qi, Wei Zhai, Dan Luo, Xiaoqin Wang, Guanghui Fu, and Bing Xiang Yang. 2024. Ai-enhanced cognitive behavioral therapy: Deep learning and large language models for extracting cognitive pathways from social media texts. *Preprint*, arXiv:2404.11449.

Tian Li, Xiang Chen, Shanghang Zhang, Zhen Dong, and Kurt Keutzer. 2020. Cross-domain sentiment classification with in-domain contrastive learning. *Preprint*, arXiv:2012.02943.

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.

Pengfei Liu, Xipeng Qiu, and Xuanjing Huang. 2017. Adversarial multi-task learning for text classification. *Preprint*, arXiv:1704.05742.

Quanyu Long, Tianze Luo, Wenya Wang, and Sinno Pan. 2022. Domain confused contrastive learning for unsupervised domain adaptation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2982–2995, Seattle, United States. Association for Computational Linguistics.

Yun Luo, Fang Guo, Zihan Liu, and Yue Zhang. 2022. Mere contrastive learning for cross-domain sentiment analysis. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 7099–7111, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Leland McInnes, John Healy, and James Melville. 2020. Umap: Uniform manifold approximation and projection for dimension reduction. *Preprint*, arXiv:1802.03426.

Quinn McNemar. 1947. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2):153–157.

Deniz Nazarova. 2023. Application of artificial intelligence in mental healthcare: Generative pre-trained transformer 3 (gpt-3) and cognitive distortions. In *Lecture Notes in Networks and Systems*, volume 813 LNNS, pages 204–219. Springer Science and Business Media Deutschland GmbH.

James Overholser and Eleanor Beale. 2023. The art and science behind socratic questioning and guided discovery: a research review. *Psychotherapy Research*, 33(7):946–956. PMID: 36878221.

Jacqueline B. Persons, Craig D. Marker, and Emily N. Bailey. 2023. Changes in affective and cognitive distortion symptoms of depression are reciprocally related during cognitive behavior therapy. *Behaviour Research and Therapy*, 166:104338.

Jonas Pfeiffer, Andreas Rücklé, Clifton Poth, Aishwarya Kamath, Ivan Vulic, Sebastian Ruder, Kyunghyun Cho, and Iryna Gurevych. 2020. Adapterhub: A framework for adapting transformers. *CoRR*, abs/2007.07779.

Barbara Plank and Gertjan Van Noord. 2011. Effective measures of domain similarity for parsing. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 1566–1576.

Alan Ramponi and Barbara Plank. 2020. Neural unsupervised domain adaptation in NLP—A survey. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6838–6855, Barcelona, Spain (Online). International Committee on Computational Linguistics.

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.

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.

Statistics Netherlands. 2024. Young adults more negative about their own health. Accessed: 14-04-2025.

Xiaomeng Wang, Dharmendra Sharma, and Dinesh Kumar. 2024. Cognitive reframing via large language models for enhanced linguistic attributes. In *The Second Tiny Papers Track at ICLR 2024*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

World Health Organization. 2024. Mental health of adolescents. Accessed: 01-05-2025.

Hongli Zhan, Allen Zheng, Yoon Kyung Lee, Jina Suh, Junyi Jessy Li, and Desmond C. Ong. 2024. Large language models are capable of offering cognitive reappraisal, if guided. *Preprint*, arXiv:2404.01288.

Jie Zhou, Junfeng Tian, Rui Wang, Yuanbin Wu, Wenming Xiao, and Liang He. 2020. SentiX: A sentiment-aware pre-trained model for cross-domain sentiment analysis. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 568–579, Barcelona, Spain (Online). International Committee on Computational Linguistics.

A Appendix

A.1 De Kindertelefoon Data

We scrape data from the Kindertelefoon forums. As the forums are moderated, it already enforces some forum rules[3] that prohibit users from sharing personally identifiable information. Nonetheless, we apply additional preprocessing steps, such as removing URLs. To further protect user privacy, we pseudonymize usernames by replacing each username with a unique identifier in the format `userXXXXXXXX`, where `XXXXXXXX` is a randomly generated eight digit number.

Subforum	# Posts
Emotionele problemen en gevoelens *(Emotional Problems and feelings)*	7524
Pesten *(Bullying)*	705
Relaties en Liefde *(Relationships and Love)*	6080
Gender & seksuele identiteit *(Gender & Sexual Identity)*	1182
Seksualiteit *(Sexuality)*	9999
Lichaam en Gezondheid *(Body and Health)*	4386
Verslaving *(Addiction)*	384
Thuis en Familie *(Home and Family)*	2576
Geweld *(Violence)*	318
Levensbeschouwing *(Philosophy of Life)*	103
Geld en Werk *(Money and Work)*	439
Internet en Mobiel *(Internet and Mobile)*	613
School en Studie *(School and Study)*	1512
Sport en Vrije Tijd *(Sport and Leisure)*	1357
Rechten en de Wet *(Rights and the Law)*	204
Succesverhalen *(Success stories)*	309
Overall	**37691**

Table 4: Distribution of forum posts in the scraped dataset across the 16 subforums from De Kindertelefoon.

A.1.1 Label Distributions

We report the label distributions for the English dataset from Shreevastava and Foltz (2021) and annotated De Kindertelefoon posts.

[3] https://forum.kindertelefoon.nl/over-de-kindertelefoon-54/forumregels-36128

Stage	Dataset	Class 0	Class 1	Total
1	EN	933	1593	2526
	KT	273	177	450
2	EN	14513	2864	17377
	KT	2411	176	2587

Table 5: Label distribution for EN and KT datasets. 0 means not distorted, 1 means distorted.

A.2 Cognitive Reframing and the ABC Model

Cognitive reframing is a core technique in CBT aimed at helping individuals replace cognitive distortions in a more balanced and constructive way (Beck, 1970). The process typically involves the following steps:

- **Identifying Distortions:** The first step is to make the person aware of their distorted thoughts, since most of the times they are automatic and slip by unnoticed. *(After a breakup, a person might think, "I'm destined to be alone, no one is ever going to love me.")*
- **Challenging these thoughts:** Through techniques like Socratic questioning, the thought is challenged to uncover the underlying core belief (Overholser and Beale, 2023). It involves asking a series of focused, open-ended questions that encourage reflection (Clark and Egan, 2015). *(The underlying core belief could be "I'm not worthy of love.")*
- **Reframe:** Once identified and challenged, negative thoughts can be replaced with more positive and constructive alternatives. *("Feeling scared about the future is understandable, but just because one relationship ended doesn't mean I'm unlovable. There are many opportunities ahead to meet someone who will appreciate and love me.")*

To aid the process of cognitive reframing, the ABC framework from Rational Emotive Behavior Therapy (REBT) (Ellis, 1957) offers a structured way to trace emotional or behavioral outcomes back to underlying beliefs (Ellis, 1991). A core component of REBT and one of the foundations of CBT (David et al., 2018a), the ABC model maps the entire situation in three points:

- **A (Adversity / Activating Event)**: The triggering event. *(For example, a breakup.)*
- **B (Belief)**: The interpretation or belief about the event, which may be rational or irra-

tional. *(The thought "No one will ever love me again")*

- **C (Consequence)**: The emotional and behavioral outcomes resulting from the belief. *(Social withdrawal, isolation and loneliness.)*

This framework highlights how irrational beliefs (B), rather than the events themselves (A), often lead to emotional distress (C), making it a useful framework for both detecting and reframing distorted thinking.

A.3 Annotation Guidelines

The following annotation guidelines were provided to annotators prior to beginning the labeling process. They outline the task objectives, definitions, and criteria used to ensure consistency during the annotation process. The definitions for the distortions are taken from Shreevastava and Foltz (2021).

```
Annotation Guide :

Your goal is to classify whether each input
    contains a distortion, and if it does, mark
    the sentence(s) that are distorted.
    Cognitive distortions are biased ways of
    thinking that negatively impact how people
    perceive themselves, others, and the world.
    These patterns of thinking are often
    irrational and can contribute to stress,
    anxiety, and low self-esteem. They involve
    misinterpretations, exaggerated negativity,
    or rigid thinking that distorts reality


1. All-or-Nothing Thinking: Viewing situations
    in black-and-white terms, without
    considering a middle ground.
Example Text: It really just occurred to me
    recently. Ive always had vague, small,
    random memories of it in my mind over the
    past few years. I knew it was my life, I
    never gave it much thought. But recently I
    started thinking about it more and I
    realized those vague memories were kind of
    all I had now.
Distorted part: But recently I started thinking
    about it more and I realized those vague
    memories were kind of all I had now.

2. Overgeneralization: Drawing broad conclusions
     from limited evidence.
Example Text: From Australia: Thank you for
    reading this. I find myself with a unique
    sort of thinking for a long time ( a few
    years now)which finds ultimate worthlessness
     in achievements in life and therefore
    experiencing significant lack of interest in
     life affairs.
Distorted part: I find myself with a unique sort
     of thinking for a long time ( a few years
    now)which finds ultimate worthlessness in
    achievements in life and therefore
    experiencing significant lack of interest in
     life affairs.

3. Mental Filter: Focusing only on negative
    details while ignoring positives.
Example Text: From Hawaii: I am in a solid
    relationship with a man who is quite a bit
    older than me. We have been together nearly
    two years but I have known him for 3: He has
    , of course, been in many other
    relationships and was even married for a
    short period a long time ago.
Distorted part: I am in a solid relationship
    with a man who is quite a bit older than me.

4. Should Statements: Rigid rules about how
    someone should behave.
Example Text: By all accounts, I should be
    highly successful. I know this because
    people who dont know me that well are always
     impressed by me. I am fairly good looking,
    have a high IQ, am witty, charming, can
    strike a conversation with anyone on
    anything and can come up with solutions fast
    .
Distorted part: By all accounts, I should be
    highly successful.

5. Labeling: Reducing someone to a single
    characteristic.
Example Text: I have been very good friends with
     my boyfriend for 15 years. We started
    dating 2 years ago. Since he was my good
    friend he knows every single detail about my
     past. I was very young and dumb and have
    done a lot sexual experiences with about 25
    -30 partners.
Distorted part: I was very young and dumb and
    have done a lot sexual experiences with
    about 25 -30 partners.

6. Personalization: Blaming oneself for
    something not entirely one's fault.
Example Text: From the USA: I have been in a
    relationship with my boyfriend for 6 years.
    I do not trust him. I caught him talking to
    another girl last year but all he says they
    did was just talk on the phone. He gets
    angry over everything. Nothing I do or say
    is ever right.
Distorted part: Nothing I do or say is ever
    right.

7. Magnification: Exaggerating the significance
    of problems or shortcomings.
Example Text: About a year ago I developed
    severe anxiety and had several panic attacks
     a day. Over time I developed more and more
    symptoms such as intrusive thoughts etc
    However after quite some time I developed
    very worrying symptoms that make me think I
    am developing schiz/psychosis.
Distorted part: About a year ago I developed
    severe anxiety and had several panic attacks
     a day. Over time I developed more and more
    symptoms such as intrusive thoughts etc
    However after quite some time I developed
    very worrying symptoms that make me think I
    am developing schiz/psychosis.

8. Emotional Reasoning: Assuming feelings
```

```
    reflect reality.
Example Text: I am currently in my second
    semester of college and have lost all of my
    motivation to keep up with my course load. I
     have lost my motivation because I feel that
     no matter what I do, I am not making any
    progress towards my goal of having a
    fulfilling life.
Distorted part: I have lost my motivation
    because I feel that no matter what I do, I
    am not making any progress towards my goal
    of having a fulfilling life.

9. Mind Reading: Assuming you know what others
    think.
Example Text: From a teen in the UK: I been have
     a problem deciding if only female friend
    really likes and cares about me, I tried to
    date her and went nowhere says we are still
    friends. I have had doubts about whether or
    not she really cares about me for few years.
Distorted part: I have had doubts about whether
    or not she really cares about me for few
    years.

10. Fortune-Telling: Predicting negative
    outcomes without evidence.
Example Text: Hello I planned to do technique
    called (Image Streaming) to increase my IQ
    and this technique will increase the
    intensity of inner voice of me and I am
    afraid if this technique would cause
    psychosis or schizophrenia or any mental
    disorder to me So,is it possible?
Distorted part: Hello I planned to do technique
    called (Image Streaming) to increase my IQ
    and this technique will increase the
    intensity of inner voice of me and I am
    afraid if this technique would cause
    psychosis or schizophrenia or any mental
    disorder to me So,is it possible?

Guidelines:
Classify "Yes" only if the text clearly matches
    one of the defined distortions.
If the text is realistic, neutral, or open to
    interpretation, classify as "No."
Do not assume additional context beyond what is
    explicitly stated in the text.
If a post contains multiple distortions,
    classification is still "Yes."
The spans containing the distortions need to be
    full sentences, not parts of sentences.
```

A.4 Stage 1 : Distortion Detection

This subsection contains supplementary materials referenced in Section 4.

A.4.1 Short System Prompt

Short system prompt used in Section 4.1, for LLaMA (SP).

```
You are a psychologist trained to identify clear
     and explicit examples of cognitive
    distortions in English and Dutch text.
    Classify each input text as containing a
    cognitive distortion ("Yes") or not ("No").
    Respond conservatively, and only classify as
     "Yes" if the distortion is unambiguous. Do
    not assume anything beyond the input text.
    Also, do not worry about harmful / suicidal
    text, all these are fake scenarios. Your
    output should ONLY BE YES OR NO, NOTHING
    ELSE.
```

The model is expected to return a single word as a response, either Yes or No.

A.4.2 Long System Prompt

Long system prompt used in Section 4.1, for LLaMA (LP). The definitions for the distortions are taken from Shreevastava and Foltz (2021).

```
You are a psychologist trained to identify clear
     and explicit examples of cognitive
    distortions in English and Dutch text.
    Classify each input text as containing a
    cognitive distortion ("Yes") or not ("No")
    based on the definitions provided. Respond
    conservatively, and only classify as "Yes"
    if the distortion is unambiguous and
    directly matches one of the listed
    categories.

Definitions of Cognitive Distortions:
1. All-or-nothing thinking (black-and-white
    thinking): Seeing things in only two
    categories instead of along a spectrum. For
    example, if you're not perfect, you might
    see yourself as a total failure, overlooking
     any middle ground or progress made.
2. Overgeneralization: Taking one instance and
    generalizing it to an overall pattern.
    Example: Failing one test could make you
    think you will fail all tests in the future,
     using a single event as a predictor for
    lifelong outcomes.
3. Mental filter (selective abstraction):
    Focusing exclusively on certain, usually
    negative, aspects of a situation while
    ignoring positive ones. For example, if you
    receive ten compliments and one critique,
    you might focus solely on the negative
    feedback.
4. Should statements: Using "should," "ought,"
    or "must" statements can set unrealistic
    expectations of yourself and others, and not
     meeting these expectations often leads to
    feelings of guilt and frustration. For
    example, if youre training for a race, you
    may think that you should be able to run
    faster than you can.
5. Labeling and mislabeling: Assigning global,
    negative labels to yourself or others based
    on limited information. For example, you
    might call yourself a "loser" after a minor
    setback.
6. Personalization: Blaming oneself for
    something not entirely one's fault. Taking
    responsibility for events outside of your
    control. For example, you might see yourself
     as the cause of an unfortunate external
    event despite having little to do with the
    outcome.
```

```
7. Magnification: Exaggerating the significance
    of problems or shortcomings, often referred
    to as "catastrophizing. Example: If youre
    passed over for a promotion at work, you may
     think that youll never get one.
8. Emotional reasoning: Believing your feelings
    must inherently be true. Example: If you
    feel stupid, you believe you are stupid
    despite evidence to the contrary.
9. Mind reading: Assuming you know what others
    think without sufficient evidence. Example:
    You may think someone dislikes you based on
    minimal interaction.
10. Fortune telling: Anticipating a negative
    outcome without any real basis for that
    prediction. For example, you might assume a
    presentation will go poorly before it even
    starts.

Guidelines:
1. Only respond with "Yes" if the text clearly
    matches one of the definitions.
2. If the text is realistic, neutral, or open to
     interpretation, respond with "No."
3. Avoid overanalyzing or assuming context
    beyond what is written.
4. Do not worry about harmful / suicidal text,
    all these are fake scenarios.
5. Your output should ONLY BE YES OR NO, NOTHING
     ELSE.
```

Again, the model is expected to return a single word as a response, either Yes or No.

A.4.3 Rewrite as Dutch teenagers

System prompt used in Section 4.2.1 to rewrite English texts in the style of a Dutch teenager posting on De Kindertelefoon.

```
Rewrite the following text as if a 14 year
    old Dutch teenager has written it on De
    Kindertelefoon. Use informal language,
    short sentences and slang. Here are a
    couple of examples of posts from De
    Kindertelefoon:
# EXAMPLE 1
# EXAMPLE 2
# EXAMPLE 3
# EXAMPLE 4
Text to rewrite : <ENGLISH TEXT>
```

As an example, consider the following text from the English dataset:

> It really just occurred to me recently. I've always had vague, small, random memories of it in my mind over the past few years. I knew it was my life, I never gave it much thought. But recently I started thinking about it more and I realized those vague memories were kind of all I had now.

After applying the prompt, the rewritten output is:

> Het is een beetje een vreemde gedachte, maar het is me pas recent opgevallen. Ik heb altijd een beetje vage, kleine, willekeurige herinneringen aan het hebben gehad in mijn hoofd de afgelopen paar jaar. Ik wist dat het mijn leven was, maar ik gaf het nooit echt veel na. Maar recentelijk ben ik er meer over gaan denken en ik realiseerde me dat die vage herinneringen eigenlijk alles wat ik nu nog over het hebben heb.

A.4.4 Empath Features

The following is the set of 68 significant Empath features used to construct the feature vector in Section 4.2.2.

```
['wedding', 'domestic_work', 'medical_emergency
', 'cold', 'hate', 'envy', 'anticipation',
'family', 'vacation', 'masculine', 'dispute
', 'nervousness', 'weakness', 'horror',
'swearing_terms', 'leisure', 'suffering',
 'royalty', 'tourism', 'kill', 'ridicule
', 'optimism', 'home', 'sexual', 'fear', '
irritability', 'driving', 'exasperation', '
internet', 'leader', 'body', 'noise', 'zest',
'confusion', 'heroic', 'celebration', 'violence
', 'neglect', 'love', 'sympathy', 'trust', '
ancient', 'deception', 'air_travel', 'toy', '
disgust', 'gain', 'youth', 'sadness', 'emotional
', 'joy', 'traveling', 'ugliness', 'lust', '
shame', 'anger', 'strength', 'power', 'party
', 'pain', 'timidity', 'negative_emotion', '
messaging', 'competing', 'friends', 'children', '
monster', 'contentment']
```

A.4.5 McNemar's Test

Since there was no clear winner in terms of performance in Table 2, we conduct pairwise McNemar's tests among the four best performing methods to evaluate whether the differences in their performances are statistically significant.

McNemar's test is a non parametric statistical test used to compare the performance of two classifiers on the same data, specifically focusing on the instances where the classifiers disagree (McNemar, 1947). It tests the null hypothesis that both models have the same error rate.

To account for multiple comparisons across the six pairwise tests, we apply Bonferroni correction (Dunn, 1961), which adjusts the significance threshold to reduce the likelihood of Type I errors. Specifically, we divide the original signifi-

Method	p value	Reject
Adapters vs DCCL Correct	0.0046	True
Adapters vs XLMR+Empath	0.3424	False
Adapters vs DCCL Naive	0.3799	False
DCCL Correct vs XLMR+Empath	0.0637	False
DCCL Correct vs DCCL Naive	0.1244	False
XLMR+Empath vs DCCL Naive	0.9110	False

Table 6: Results of the pairwise McNemar's test. Reject=True means you reject the null hypothesis, which states that the two models perform equally (no significant difference between them).

cance level ($\alpha = 0.05$) by the number of comparisons ($k = 6$), resulting in an adjusted threshold of $\alpha' = \frac{0.05}{6} \approx 0.0083$. The results are in Table 6.

A.4.6 Analysis of Classifier Outputs

We compare the predictions of the 4 best performing classifiers from Section 4 for the subset of data that was annotated by both the annotators.

In Table 7, we see some interesting patterns. XLMR+Empath predicts a text as distorted 49% of the time, showing a nearly balanced prediction ratio (51/49). In contrast, Adapters (24%), DCCL (23%), and Naïve DCCL (33%) show a clear bias toward the non distorted class. This suggests that XLMR+Empath is more liberal in flagging positive cases, which may be beneficial in high recall applications, though potentially at the cost of precision.

Across all models, the number of "Not Confusing" cases are higher than "Confusing" ones. This indicates that when models fail, they often do so on examples where human annotators agreed independently. This pattern suggests a model 'blind spot' on straightforward cases. However, there needs to be a detailed analysis done to see what is causing it.

There is an asymmetry in model disagreements:

- For **Adapters**, 87% of disagreements are false negatives (predicting not distorted when both annotators labeled distorted).
- For **DCCL**, the false negative rate among disagreements is even higher at 92%.
- For **Naïve DCCL**, this drops slightly to 78% (18 out of 23).
- In contrast, **XLM-R Empath** shows a reverse trend: 14 out of 24 disagreements (58%) are false positives (predicting distorted when annotators labeled not distorted).

These patterns reveal asymmetric model behaviour. XLMR+Empath is more prone to false alarms, whereas the other models tend to underpredict positives, suggesting a more conservative outlook. There needs to be a careful consideration of the tradeoff between false positives and false negatives when selecting a model for deployment.

A.5 Stage 2 : Identifying the distorted span

This subsection contains supplementary materials referenced in Section 5.

A.5.1 Prompting LLaMA - Unstructured Output

Modelfile used to create the LLaMA model used in Section 5.1.1, for the unstructured output. A Modelfile in Ollama is a configuration file that defines how a custom model should be built, including its base model and behavior. The common phrases per distortion are taken from Bathina et al. (2020).

```
FROM finalend/hermes-3-llama-3.1:latest

PARAMETER temperature 0

SYSTEM """
You are a very smart cognitive behavioral
    therapist, who is trained to identify
    occurences of cognitive distortions in text.
     These are some common phrases in English
    associated with a distortion.

All or nothing thinking.
only, every, everyone, everybody, everything,
    everywhere, always, perfect, the best, all,
not a single, no one, nobody, nothing, nowhere,
    never, worthless, the worst, neither, nor,
either or, black or white, ever

Overgeneralizing
all ofthe time, all ofthem, all the time, always
     happens, always like, happens every time,
completely, no one ever, nobody ever, every
    single one ofthem, every single one ofyou, I
always, you always, he always, she always, they
    always, I am always, you are always, he
is always, she is always, they are always

Mental Filtering
I see only, all I see, all I can see, can only
    think, nothing good, nothing right,
    completely
bad, completely wrong, only the bad, only the
    worst, ifI just, ifI only, if it just, if it
     only
```

Scenario	DCCL	Naïve DCCL	Adapters	XLMR+Empath
Predictions (0, 1)	77, 23	67, 33	76, 24	51, 49
Model agrees with annotators	66 (0.66)	71 (0.71)	63 (0.63)	70 (0.70)
Model agrees with annotators (Prediction=1, True=1)	20 (0.31)	28 (0.39)	19 (0.30)	36 (0.51)
Model agrees with annotators (Prediction=0, True=0)	46 (0.69)	43 (0.61)	44 (0.70)	34 (0.49)
Model disagrees with annotators	28 (0.28)	23 (0.23)	31 (0.31)	24 (0.24)
Model disagrees with annotators (Prediction=0, True=1)	26 (0.92)	18 (0.78)	27 (0.87)	10 (0.42)
Model disagrees with annotators (Prediction=1, True=0)	2 (0.08)	5 (0.22)	4 (0.13)	14 (0.58)
Confusing	6 (0.21)	5 (0.22)	8 (0.25)	5 (0.20)
Not Confusing	22 (0.79)	18 (0.78)	23 (0.75)	19 (0.80)

Table 7: Model agreement and disagreement scenarios across different methods. The first row ("Predictions") shows the number of instances predicted as not distorted (class 0) and distorted (class 1), respectively. Percentages are shown in parentheses. For disagreement cases between the model and annotators, we further categorize them as Confusing if the annotators initially disagreed before deliberation, and Not Confusing if they had already agreed.

```
Should Statements
blame me, I caused, I feel responsible, all my doing, all my fault, my bad, my responsibility
should, ought, must, have to, has to

Labeling
I am a, he is a, she is a, they are a, it is a, that is a, sucks at, suck at, I never, he never,
she never, you never, we never, they never, I am an, he is an, she is an, they are an, it is an, that is an, a burden, a complete, a completely, a huge, a loser, a major, a total, a
totally, a weak, an absolute, an utter, a bad, a broken, a damaged, a helpless, a hopeless, an incompetent, a toxic, an ugly, an undesirable, an unlovable, a worthless, a horrible, a terrible

Personalization
all me, all my, because I, because my, because of my, because of me, I am responsible,

Magnication
worst, best, not important, not count, not matter, no matter, the only thing, the one thing

Emotional Reasoning
but I feel, since I feel, because I feel, but it feels, since it feels, because it feels, still feels

Mindreading
everyone believes, everyone knows, everyone thinks, everyone will believe, everyone will know, everyone will think, nobody believes, nobody knows, nobody thinks, nobody will believe, nobody will know, nobody will think, he believes, he knows, he thinks, he does not believe, he does not know, he does not think, he will believe, he will know, he will think, he will not believe, he will not know, he will not think, she believes, she knows, she thinks,
she does not believe, she does not know, she does not think, she will believe, she will know,
she will think, she will not believe, she will not know, she will not think, they believe, they
know, they think, they do not believe, they do not know, they do not think, they will believe,
they will know, they will think, they will not believe, they will not know, they will not think,
we believe, we know, we think, we do not believe, we do not know, we do not think, we
will believe, we will know, we will think, we will not believe, we will not know, we will
not think, you believe, you know, you think, you do not believe, you do not know, you do
not think, you will believe, you will know, you will think, you will not believe, you will not
know, you will not think

Fortune-telling
I will not, we will not , you will not, they will not, it will not, that will not, he will not, she
will not

If these occur in a sentence, it is likely to be distorted
```

```
Here are some examples :
Input Text: "From Australia: Thank you for
    reading this. I find myself with a unique
    sort of thinking for a long time (a few
    years now) which finds ultimate
    worthlessness in achievements in life and
    therefore experiencing significant lack of
    interest in life affairs."
Distortion: "I find myself with a unique sort of
     thinking for a long time (a few years now)
    which finds ultimate worthlessness in
    achievements in life and therefore
    experiencing significant lack of interest in
     life affairs."

Input Text: "From Hawaii: I am in a solid
    relationship with a man who is quite a bit
    older than me. We have been together nearly
    two years but I have known him for 3: He has
    , of course, been in many other
    relationships and was even married for a
    short period a long time ago."
Distortion: "I am in a solid relationship with a
     man who is quite a bit older than me."

Input Text: "By all accounts, I should be highly
     successful. I know this because people who
    dont know me that well are always impressed
    by me. I am fairly good looking, have a high
     IQ, am witty, charming, can strike a
    conversation with anyone on anything and can
     come up with solutions fast."
Distortion: "By all accounts, I should be highly
     successful."

Input Text: "I have been very good friends with
    my boyfriend for 15 years. We started dating
     2 years ago. Since he was my good friend he
     knows every single detail about my past. I
    was very young and dumb and have done a lot
    sexual experiences with about 25 -30
    partners."
Distortion: "I was very young and dumb and have
    done a lot sexual experiences with about 25
    -30 partners."



These are your guidelines:
1. Extract ONLY COMPLETE SENTENCES from the
    input text that contain cognitive
    distortions
2. Never modify wording, truncate, or paraphrase
     - preserve original text exactly
3. If multiple sentences: maintain original
    order and COMBINE INTO ONE PARAGRAPH
4. NEVER use lists/bullets/numbering - only
    continuous text
5. Output MUST follow this format: Distortion:
    followed by the sentence / sentences with NO
     other text
6. There is atleast one distorted sentence in
    the input text
"""
```

A.5.2 Prompting LLaMA - Structured Output

Prompt used for the structured output configuration in Section 5.1.1.

```
template = '''
### Task:
You are a psychologist, tasked with finding
    distorted sentences in input text. You
    should follow these rules STRICTLY in order:
1. I will provide you with an input text, split
    into sentences. These sentences follow each
    other (if you concatenate them together they
     make up the original input text). For each
    of the sentence in the list, determine if
    that is distorted.
2. The output should be a list of zeros and ones,
     where ones represent distorted sentences,
    and zeros means not distorted.
3. The length of the output array SHOULD BE
    EQUAL TO THE NUMBER OF INPUT SENTENCES.

### Examples:
Input Text: "
Sentence 1 : From Australia: Thank you for
    reading this.
Sentence 2 : I find myself with a unique sort of
     thinking for a long time (a few years now)
    which finds ultimate worthlessness in
    achievements in life and therefore
    experiencing significant lack of interest in
     life affairs."
Distorted sentence(s): "I find myself with a
    unique sort of thinking for a long time (a
    few years now) which finds ultimate
    worthlessness in achievements in life and
    therefore experiencing significant lack of
    interest in life affairs."
Output: [0, 1]

Input Text: "
Sentence 1 : By all accounts, I should be highly
     successful.
Sentence 2 : I know this because people who dont
     know me that well are always impressed by
    me.
Sentence 3 : I am fairly good looking, have a
    high IQ, am witty, charming, can strike a
    conversation with anyone on anything and can
     come up with solutions fast."
Distorted sentence(s): "By all accounts, I
    should be highly successful."
Output: [1, 0, 0]

Input Text: "
Sentence 1 : I have been very good friends with
    my boyfriend for 15 years.
Sentence 2 : We started dating 2 years ago.
Sentence 3 : Since he was my good friend he
    knows every single detail about my past.
Sentence 4 : I was very young and dumb and have
    done a lot sexual experiences with about 25
    -30 partners."
Distorted sentence(s): "I was very young and
    dumb and have done a lot sexual experiences
    with about 25 -30 partners."
Output: [0, 0, 0, 1]

Now, give me the output for this piece of text.
    ONLY OUTPUT THE ARRAY, NOTHING ELSE.
Text:
{input_text}
'''
```

B Experimental and Hyperparameter Details

B.1 Experimental Details

We outline the experimental setup for both stages. All our code is based on the Huggingface library (Wolf et al., 2019). For XLM-RoBERTa based methods, we use *xlm-roberta-base* as the encoder. For LLaMA with a classification head, we use *meta-llama/Llama-3.1-8B*. In Stage 1, prompting and instruction tuning on LLaMA is conducted using Unsloth (Daniel Han and team, 2023), specifically with the *unsloth/Meta-Llama-3.1-8B-Instruct-bnb-4bit* model. In Stage 2, prompting is done using Ollama, on *finalend/hermes-3-llama-3.1:latest*. All experiments are performed on a NVIDIA A40 GPU.

B.2 Hyperparameter Details

Below are the hyperparameters for the models used in Section 4 and Section 5. If a hyperparameter is not mentioned, default values from the HuggingFace Trainer or Unsloth notebooks are used.

Method	LR	Epochs	Weight Decay
Adapters	1×10^{-4}	6	–
Finetuned XLMR	5×10^{-5}	6	–
XLMR+Empath	2×10^{-5}	3	0.01
DCCL (TL1)	1×10^{-5}	3	0.01
DCCL (TL2)	2×10^{-5}	2	0.01
Naive DCCL (TL1)	1×10^{-5}	3	0.01
Naive DCCL (TL2)	2×10^{-5}	3	0.01

Table 8: Hyperparameters for all models used in Stage 1 and Stage 2 experiments. Second column (LR) means the learning rate. For DCCL and Naive DCCL, TL1 means the first training loop, and TL2 is the second training loop.

C Extra Tables

C.1 Stage 1 - Non distorted Class

Table 9 reports the Precision, Recall and F1 scores for the non distorted class of the Section 4 models.

C.2 Stage 1 - Distorted Class

Table 10 reports the Precision, Recall and F1 scores for the distorted class of Section 4 models.

Train Data	Method	Precision	Recall	F1
None	LLaMA SP	0.64 ± 0.11	0.19 ± 0.05	0.29 ± 0.08
None	LLaMA LP	0.65 ± 0.06	0.29 ± 0.04	0.39 ± 0.04
None	Random Baseline	0.63 ± 0.06	0.51 ± 0.07	0.56 ± 0.05
EN	Adapters	0.93 ± 0.03	0.36 ± 0.06	0.51 ± 0.06
EN	LLaMA Finetuning	0.62 ± 0.04	0.77 ± 0.28	0.66 ± 0.15
EN	XLMR Finetuning	0.89 ± 0.09	0.33 ± 0.11	0.47 ± 0.11
EN	LLaMA Instruction Tuning	0.67 ± 0.12	0.38 ± 0.06	0.48 ± 0.06
Rewritten EN	XLMR Finetuning	0.91 ± 0.09	0.28 ± 0.14	0.40 ± 0.15
EN+KT	Adapters	0.72 ± 0.06	0.76 ± 0.07	0.74 ± 0.03
EN+KT	LLaMA Finetuning	0.65 ± 0.04	0.82 ± 0.11	0.72 ± 0.04
EN+KT	XLMR Finetuning	0.51 ± 0.29	0.75 ± 0.43	0.60 ± 0.34
EN+KT	LLaMA Instruction Tuning	0.71 ± 0.04	0.71 ± 0.02	0.71 ± 0.03
EN+KT	XLMR+Empath	0.75 ± 0.07	0.77 ± 0.09	0.76 ± 0.05
EN+KT	Naive DCCL	0.75 ± 0.10	0.79 ± 0.15	0.76 ± 0.06
EN+KT	DCCL	0.78 ± 0.06	0.74 ± 0.07	0.76 ± 0.04

Table 9: Precision, Recall, and F1 Scores for the non distorted class for Stage 1 Methods. "Train Data" indicates the dataset used for training: None (zero shot), EN (original English dataset), Rewritten EN (English data rewritten in the style of De Kindertelefoon), and KT (De Kindertelefoon). For LLaMA, SP denotes Short Prompt and LP denotes Long Prompt.

Train Data	Method	Precision	Recall	F1
None	LLaMA SP	0.40 ± 0.03	0.85 ± 0.03	0.55 ± 0.03
None	LLaMA LP	0.42 ± 0.04	0.77 ± 0.08	0.55 ± 0.05
None	Random Baseline	0.41 ± 0.05	0.54 ± 0.10	0.46 ± 0.05
EN	Adapters	0.48 ± 0.02	0.96 ± 0.02	0.64 ± 0.02
EN	LLaMA Finetuning	0.37 ± 0.26	0.27 ± 0.28	0.28 ± 0.21
EN	XLMR Finetuning	0.48 ± 0.05	0.93 ± 0.06	0.63 ± 0.03
EN	LLaMA Instruction Tuning	0.42 ± 0.06	0.70 ± 0.16	0.52 ± 0.09
Rewritten EN	XLMR Finetuning	0.46 ± 0.03	0.94 ± 0.07	0.61 ± 0.02
EN+KT	Adapters	0.60 ± 0.03	0.54 ± 0.13	0.56 ± 0.09
EN+KT	LLaMA Finetuning	0.54 ± 0.09	0.30 ± 0.11	0.37 ± 0.07
EN+KT	XLMR Finetuning	0.40 ± 0.42	0.32 ± 0.45	0.24 ± 0.30
EN+KT	LLaMA Instruction Tuning	0.54 ± 0.06	0.54 ± 0.08	0.54 ± 0.07
EN+KT	XLMR+Empath	0.64 ± 0.08	0.60 ± 0.14	0.61 ± 0.09
EN+KT	Naive DCCL	0.71 ± 0.17	0.56 ± 0.30	0.54 ± 0.27
EN+KT	DCCL	0.63 ± 0.04	0.67 ± 0.11	0.65 ± 0.05

Table 10: Precision, Recall, and F1 Scores for the distorted class for Stage 1 Methods. "Train Data" indicates the dataset used for training: None (zero shot), EN (original English dataset), Rewritten EN (English data rewritten in the style of De Kindertelefoon), and KT (De Kindertelefoon). For LLaMA, SP denotes Short Prompt and LP denotes Long Prompt.

C.3 Results for Encoder + Linear Classifier

Weighted Precision, Recall and F1 score and Precision, Recall and F1 score for the Encoder + Linear Classifier configuration in Section 5.

Method	P (0)	R (0)	F1 (0)	P (1)	R (1)	F1 (1)	P	R	F1
BC	0.94 ± 0.00	0.98 ± 0.01	0.95 ± 0.01	0.24 ± 0.14	0.13 ± 0.07	0.17 ± 0.10	0.88 ± 0.01	0.92 ± 0.01	0.90 ± 0.00
BM	0.91 ± 0.04	0.99 ± 0.01	0.96 ± 0.00	0.16 ± 0.17	0.04 ± 0.05	0.07 ± 0.07	0.88 ± 0.02	0.92 ± 0.01	0.89 ± 0.00
DC	0.93 ± 0.01	0.98 ± 0.01	0.96 ± 0.00	0.22 ± 0.16	0.10 ± 0.09	0.14 ± 0.11	0.88 ± 0.02	0.92 ± 0.01	0.90 ± 0.01
DM	0.93 ± 0.00	0.99 ± 0.01	0.96 ± 0.00	0.13 ± 0.19	0.06 ± 0.10	0.08 ± 0.12	0.87 ± 0.02	0.93 ± 0.01	0.90 ± 0.01
NC	0.93 ± 0.01	0.98 ± 0.01	0.96 ± 0.00	0.29 ± 0.18	0.15 ± 0.10	0.19 ± 0.13	0.89 ± 0.02	0.92 ± 0.01	0.90 ± 0.01
NM	0.93 ± 0.01	0.99 ± 0.01	0.96 ± 0.00	0.13 ± 0.18	0.06 ± 0.08	0.08 ± 0.10	0.87 ± 0.02	0.92 ± 0.01	0.90 ± 0.01

Table 11: Precision, Recall, and F1 scores for both classes, along with weighted Precision, Recall, and F1, for the Encoder + Linear Classifier setup in Section 5.1.2. Each method is denoted by two letters: the first letter represents the encoder (B = XLM-RoBERTa, D = DCCL, N = Naive DCCL), and the second letter denotes the embedding type (C = [CLS] token, M = mean pooled). 0 represents the non distorted class, and 1 represents the distorted class.

C.4 Results for Encoder + SVM

Results for all possible combinations for Encoder + SVM in Stage 5.

Method	P (0)	R (0)	F1 (0)	P (1)	R (1)	F1 (1)	P	R	F1
BLC	0.65 ± 0.10	0.68 ± 0.12	0.66 ± 0.10	0.67 ± 0.10	0.64 ± 0.11	0.65 ± 0.10	0.66 ± 0.09	0.66 ± 0.10	0.66 ± 0.10
BPC	0.79 ± 0.03	0.56 ± 0.09	0.65 ± 0.07	0.66 ± 0.05	0.85 ± 0.03	0.74 ± 0.03	0.73 ± 0.02	0.70 ± 0.04	0.70 ± 0.05
BRC	0.73 ± 0.06	0.68 ± 0.06	0.70 ± 0.02	0.70 ± 0.04	0.74 ± 0.08	0.72 ± 0.03	0.72 ± 0.02	0.71 ± 0.02	0.71 ± 0.02
BSC	0.66 ± 0.07	0.66 ± 0.06	0.66 ± 0.06	0.66 ± 0.05	0.66 ± 0.06	0.66 ± 0.06	0.66 ± 0.06	0.66 ± 0.06	0.66 ± 0.06
BLM	0.67 ± 0.04	0.69 ± 0.06	0.68 ± 0.04	0.68 ± 0.05	0.66 ± 0.04	0.67 ± 0.04	0.68 ± 0.04	0.68 ± 0.04	0.68 ± 0.04
BPM	0.80 ± 0.07	0.56 ± 0.06	0.66 ± 0.07	0.66 ± 0.06	0.86 ± 0.06	0.75 ± 0.06	0.73 ± 0.07	0.71 ± 0.06	0.70 ± 0.07
BRM	0.77 ± 0.06	0.72 ± 0.10	0.74 ± 0.07	0.74 ± 0.09	0.78 ± 0.08	0.76 ± 0.07	0.76 ± 0.06	0.75 ± 0.07	0.75 ± 0.07
BSM	0.72 ± 0.07	0.74 ± 0.08	0.73 ± 0.06	0.73 ± 0.09	0.71 ± 0.09	0.72 ± 0.08	0.73 ± 0.07	0.72 ± 0.07	0.72 ± 0.07
DLC	0.67 ± 0.07	0.67 ± 0.06	0.67 ± 0.04	0.67 ± 0.05	0.67 ± 0.10	0.66 ± 0.06	0.67 ± 0.05	0.67 ± 0.05	0.66 ± 0.05
DPC	0.52 ± 0.02	0.97 ± 0.02	0.68 ± 0.01	0.72 ± 0.21	0.10 ± 0.07	0.16 ± 0.11	0.62 ± 0.11	0.53 ± 0.02	0.42 ± 0.05
DRC	0.61 ± 0.05	0.84 ± 0.01	0.71 ± 0.03	0.74 ± 0.02	0.46 ± 0.09	0.57 ± 0.07	0.68 ± 0.03	0.65 ± 0.04	0.64 ± 0.05
DSC	0.54 ± 0.05	0.56 ± 0.07	0.55 ± 0.03	0.54 ± 0.04	0.52 ± 0.13	0.52 ± 0.08	0.54 ± 0.04	0.54 ± 0.04	0.54 ± 0.04
DLM	0.66 ± 0.08	0.61 ± 0.06	0.63 ± 0.07	0.64 ± 0.06	0.68 ± 0.08	0.66 ± 0.06	0.65 ± 0.06	0.64 ± 0.06	0.64 ± 0.06
DPM	0.72 ± 0.02	0.70 ± 0.08	0.70 ± 0.05	0.71 ± 0.06	0.73 ± 0.03	0.72 ± 0.04	0.71 ± 0.04	0.71 ± 0.04	0.71 ± 0.04
DRM	0.69 ± 0.03	0.70 ± 0.08	0.69 ± 0.05	0.70 ± 0.07	0.69 ± 0.04	0.69 ± 0.04	0.70 ± 0.05	0.69 ± 0.05	0.69 ± 0.05
DSM	0.68 ± 0.06	0.77 ± 0.06	0.72 ± 0.04	0.73 ± 0.05	0.63 ± 0.09	0.67 ± 0.06	0.71 ± 0.05	0.70 ± 0.05	0.70 ± 0.05
NLC	0.61 ± 0.09	0.62 ± 0.12	0.61 ± 0.08	0.62 ± 0.10	0.60 ± 0.12	0.61 ± 0.08	0.62 ± 0.08	0.61 ± 0.08	0.61 ± 0.08
NPC	0.65 ± 0.05	0.44 ± 0.09	0.52 ± 0.07	0.58 ± 0.06	0.76 ± 0.04	0.66 ± 0.04	0.61 ± 0.05	0.60 ± 0.05	0.59 ± 0.06
NRC	0.63 ± 0.05	0.66 ± 0.11	0.64 ± 0.07	0.65 ± 0.09	0.62 ± 0.08	0.63 ± 0.06	0.64 ± 0.06	0.64 ± 0.06	0.64 ± 0.06
NSC	0.61 ± 0.13	0.57 ± 0.10	0.59 ± 0.11	0.59 ± 0.07	0.63 ± 0.13	0.61 ± 0.09	0.60 ± 0.10	0.60 ± 0.09	0.60 ± 0.09
NLM	0.64 ± 0.09	0.66 ± 0.10	0.65 ± 0.08	0.65 ± 0.07	0.62 ± 0.10	0.63 ± 0.06	0.65 ± 0.07	0.64 ± 0.06	0.64 ± 0.06
NPM	0.73 ± 0.09	0.43 ± 0.07	0.54 ± 0.05	0.60 ± 0.04	0.84 ± 0.06	0.69 ± 0.03	0.67 ± 0.05	0.63 ± 0.04	0.62 ± 0.04
NRM	0.67 ± 0.06	0.66 ± 0.07	0.66 ± 0.05	0.67 ± 0.06	0.68 ± 0.08	0.67 ± 0.06	0.67 ± 0.05	0.67 ± 0.05	0.67 ± 0.05
NSM	0.61 ± 0.09	0.60 ± 0.08	0.60 ± 0.07	0.60 ± 0.07	0.60 ± 0.14	0.59 ± 0.09	0.60 ± 0.07	0.60 ± 0.07	0.60 ± 0.07

Table 12: Precision, Recall, and F1 scores for both classes, along with weighted Precision, Recall, and F1, for the Encoder + SVM setup in Section 5.1.2. Each method is denoted by three letters: the first letter represents the encoder (B = XLM-RoBERTa, D = DCCL, N = Naive DCCL), the second letter the kernel (L = Linear, P = Poly, R = RBF, S = Sigmoid), and the third letter denotes the embedding type (C = [CLS] token, M = mean pooled). 0 represents the non distorted class, and 1 represents the distorted class.