# Comparative Analysis of the Efficacy in the Classification of Cognitive Distortions Using LLMs

Aaron Pico[a], Joaquin Taverner[b], Emilio Vivancos[c] and Ana Garcia-Fornes[d]
*Valencian Research Institute for A fi cnica de València, Valencia, Spain*
*{apicpas, joataap, vivancos, agarcia}@upv.es*

Keywords: Cognitive Distortion, Cognitive Distortion Recognition, Large Language Model, Affective Computing, Mental Health.

Abstract: This paper explores the application of Large Language Models (LLMs) for the classification of cognitive distortions in humans. This is important for detecting irrational thought patterns that may negatively influence people's emotional state. To achieve this, we evaluated a range of open-source LLMs with varying sizes and architectures to assess their effectiveness in the task. The results show promising results of the recognition capabilities of these models, particularly given that none of them were specifically fine-tuned for this task and were solely guided by a structured prompt. The results allow us to see a trend where larger models generally outperform their smaller counterparts in this task. However, architecture and training strategies are also important factors, as some smaller models achieve performance levels comparable to or exceeding larger ones. This study has also allowed us to see the limitations in this field: the subjectivity factor that may exist in the annotations of cognitive distortions due to overlapping categories. This ambiguity impacts both human agreement and model performance. Therefore, future work includes fine-tuning LLMs specifically for this task and improving the quality of the dataset to improve performance and address ambiguity.

## 1 INTRODUCTION

Today's lifestyle, socioeconomic problems, stress and unwanted loneliness have led to an increase in mental health problems that exceed the capacity of health systems. Artificial intelligence can help to solve this problem contributing to improve the mental health by providing support to therapists and professionals, as well as facilitating the personalization and continuous monitoring of mental health (Uban et al., 2021).

According to the APA online dictionary, a cognitive distortion is a "faulty or inaccurate thinking, perception, or belief" (APA, 2024). Cognitive distortions are non-rational thoughts that alter our perception of events in our environment, and consequently our emotional state and behaviour (Yurica and DiTomasso, 2005). Cognitive distortions reflect an internal biases view of reality which increases the likelihood of mental illnesses, such as depression (Beck, 1963; Beck, 1964; Blake et al., 2016; Mahali et al., 2020; Jager-Hyman et al., 2014).

It is normal for cognitive distortions to appear occasionally in our lives. They often appear as a consequence of periods of high stress or traumatic events and can occur more or less frequently in all people. The problem appears when this way of interpreting reality does not disappear and becomes a common bias in the perception of our environment. When a person reasons using a cognitive distortion, he or she is hardly aware of the bias that is occurring in his or her interpretation of reality. This is one of the reasons why a person is often unaware of the occurrence of a cognitive distortion (Beck, 1995).

The difficulty of detecting cognitive distortions also extends to professionals in therapy (Fortune and Goodie, 2012). Therapies such as cognitive behavioural therapy (Day, 2017) makes patients compare their thoughts, feelings and behaviours with the sources that provoke them, but it is necessary to detect this reasoning influenced by cognitive distortions.

When a cognitive distortion is detected during a conversation or in a written text, in order for the person to be aware of his cognitive bias, it is important to be able to specify how it has occurred. For this reason, the recognition of cognitive distortions is es-

[a] https://orcid.org/0000-0002-5612-8033
[b] https://orcid.org/0000-0002-5163-5335
[c] https://orcid.org/0000-0002-0213-0234
[d] https://orcid.org/0000-0003-4482-8793

Pico, A., Taverner, J., Vivancos, E. and Garcia-Fornes, A.
Comparative Analysis of the Efficacy in the Classification of Cognitive Distortions Using LLMs.
DOI: 10.5220/0013399200003890
Paper published under CC license (CC BY-NC-ND 4.0)
In *Proceedings of the 17th International Conference on Agents and Artificial Intelligence (ICAART 2025) - Volume 1*, pages 957-965
ISBN: 978-989-758-737-5; ISSN: 2184-433X
Proceedings Copyright © 2025 by SCITEPRESS – Science and Technology Publications, Lda.

sential for a correct diagnosis, but also to personalize the therapy according to the distortion detected. One of the most commonly used classifications in psychiatry divides the cognitive distortions into different categories (Hossain, 2009) each of which represents a distinct pattern of irrational thinking that influences how people interpret and respond to their environment. However, it is important to note that even the most experienced psychologists often hesitate to assign a cognitive distortion to one of these categories.

An experienced therapist may detect more cognitive distortions in his or her patient than an inexperienced one, but there will still be distortions not recognised by the therapist. Moreover, the therapist will only be able to recognise the distortions that appear during his or her conversation with the patient. In this type of scenario, a cognitive distortion recognition system can serve as an aid for the therapist during his or her sessions with patients, but also as a tool to detect distortions that occur without the presence of the therapist. The ability to analyze natural language makes Large Language Models (LLMs) a powerful tool to aid in the diagnosis and treatment of patients. Specifically, LLMs can be a very useful tool due to their ability to analyse natural language (Annepaka and Pakray, 2024). Moreover, the multimodal capabilities of LLMs could allow them for the detection of cognitive distortions in speech or writing. For example, a multimodal LLM could recognise cognitive distortions by analysing the audio of conversations in which the patient participates, or texts written by the patient on their social networks. But for a cognitive distortion recogniser to work in any of these circumstances, it needs to be deployed on a mobile device with less computational capacity than the large computing clusters where LLMs are usually run.

This raises the questions: Are LLMs capable of meeting these requirements? And, could a lightweight version fulfill this task? Addressing these questions, in this study we aim to evaluate whether LLMs of different families and sizes are currently capable of performing this task as hypothesized, or if there are indications that they might achieve this with future research and advancements. To this end, we are going to perform a cognitive distortion detection test on a corpus derived from the Therapist Q&A dataset, which includes annotated interactions between patients and therapists. This dataset is particularly appropriate for our study as it aligns closely with the real-world challenge of identifying cognitive distortions in patient-therapist conversations.

## 2 RELATED WORK

Advances in natural language processing (NLP) and machine learning let us account for cognitive distortions in texts. Prior research has explored the utility of machine learning in identifying cognitive distortions in mental health texts (Shickel et al., 2020; Simms et al., 2017), social media (Alhaj et al., 2022), journaling texts (Mostafa et al., 2021), and in medical dialogues between physicians and patients (Shreevastava and Foltz, 2021; Tauscher et al., 2023; Ding et al., 2022). Although previous studies into cognitive distortions have yielded promising results with regard to their detection (binary classification), they have encountered less favourable results when attempting to categorize multi-class cognitive distortions. Few annotated datasets on cognitive distortions exist, and researchers frequently create their own tailored to their needs and cultural context.

In contrast, the use of LLMs in classifying cognitive distortions remains a relatively unexplored area. For example, the study in (Wang et al., 2023) compares the performance of fine-tuned pre-trained models with ChatGPT in few-shot and zero-shot learning scenarios. The authors introduce the C2D2 dataset, a Chinese dataset designed to address the lack of research resources on cognitive distortions, covering seven classes: All-or-nothing thinking, Emotional reasoning, Fortune-telling, Labeling, Mind reading, Overgeneralization, and Personalization. They fine-tune various Chinese pre-trained language models (BERT, RoBERTa, XLNet, and Electra) and evaluate ChatGPT's performance in these scenarios. Results show that while the pre-trained models performed well, ChatGPT's performance did not match that of the fine-tuned models, even in the few-shot learning setting. Another interesting approach can be found in (Shickel et al., 2020). The authors present a machine learning framework for detecting and classifying 15 cognitive distortions using two datasets: CrowdDist (from crowdsourcing) and MH (from a therapy program). CrowdDist contains 7,666 text responses from 1,788 individuals sourced via Mechanical Turk, where workers provided personal examples of distorted thinking. MH consists of 1,164 annotated journal entries from TAO Connect2, an online therapy service for college students. After testing various algorithms, logistic regression emerged as the best model. For CrowdDist, the model achieved a weighted F1 score of 0.68. The MH model failed to predict seven distortions, highlighting challenges in smaller, unbalanced datasets. Random chance accuracy was 0.06.

In (Simms et al., 2017), several personal blogs were collected from the Tumblr API, labeled, and then analyzed with the software Linguistic Inquiry and Word Count (LIWC). Of the 459 posts, 207 (45.1%) were labeled as distorted and 252 (54.9%) as undistorted. The LIWC and hand-labeling yielded a vector for each post. These 459 vectors constituted the training data for the machine learning approach. The best results were obtained with a combination of RELIEF (Kira and Rendell, 1992; Kononenko, 1994) and logistic regression. The findings show that it is possible to detect cognitive distortions automatically from personal blogs with relatively good accuracy (73.0%) and a false negative rate (30.4%).

The work presented in (Alhaj et al., 2022) introduces a machine learning approach to classify five cognitive distortions (Inflexibility, Overgeneralization, Labeling, Emotional Reasoning, Catastrophizing) in Arabic Twitter content. The authors enhance classification by leveraging a transformer-based topic modeling algorithm (BERTopic) built on AraBERT. The dataset includes 9,250 annotated texts (6,940 for training and 2,310 for testing). Multiple classifiers were tested, including decision trees, k-nearest neighbors, support vector machines, random forest, extreme gradient boosting, stacking, and bagging. Results demonstrated that the enriched features from topic modeling significantly improved classifier performance.

The study in (Shreevastava and Foltz, 2021) considers ten classes of cognitive distortions: Emotional reasoning, Overgeneralization, Mental filter, Should statements, All-or-nothing thinking, Mind reading, Fortune-telling, Magnification, Personalization, and Labeling. The “Therapist Q&A” dataset was procured from the Kaggle crowdsourced repository (Section 3). The study compares five algorithms (logistic regression, support vector machines, decision trees, k-nearest neighbors with k=15, and multi-layer perceptron) in detecting cognitive distortion. SVM with pre-trained S-BERT embeddings had the best results with an F1 score of 0.79. However, the detection of the specific type of distortion yielded less favorable results. None of the algorithms achieved a weighted F1 score above 0.30.

Finally, in (Tauscher et al., 2023), the authors explored the potential of NLP methods to detect and classify cognitive distortions in text messages between clinicians and individuals with serious mental illness. The goal was to assess if NLP methods could perform as well as clinically trained human raters. Data from 39 clients diagnosed with various mental health disorders was collected between 2017 and 2019. Five cognitive distortion types were annotated by clinically trained raters, with messages often labelled with multiple distortions. Cohen’s kappa for annotator agreement was 0.51. The study used Logistic Regression (LR), Support Vector Machine (SVM), and BERT models for multi-label binary classification. BERT outperformed LR and SVM, achieving similar performance to clinical experts for most distortion types, except for “should-statements” and “overgeneralization”. The study was further improved by (Ding et al., 2022), which addressed poor classification performance for less frequent distortions by using data augmentation and the domain-specific pretrained model, MentalBERT. MentalBERT delivered the best results, particularly for dominant distortion classes, though data augmentation methods varied in effectiveness depending on distortion frequency.

As previously highlighted, there has been limited investigation into the use of LLMs for cognitive distortion classification. While the study by (Wang et al., 2023) offers an initial comparison between ChatGPT and more traditional fine-tuned models, it is focused on a single proprietary text-generating model. In our work, we extend this perspective by evaluating multiple LLMs from different families, architectures, and parameter sizes. Moreover, we include open source or freely available models that can be used on our own hardware. In this way we can explore the differences between LLMs and emphasize their accessibility and potential for practical applications in clinical and research settings.

# 3 LLMs FOR DETECT COGNITIVE DISTORTIONS

The advancements in LLMs have opened new possibilities for complex natural language processing tasks, including those tasks relevant to mental health and cognitive analysis. The following sections describe the methodology applied to evaluate the performance of selected LLMs in a cognitive distortions recognition task.

## 3.1 Methodology

### 3.1.1 Cognitive Distortions Recognition Task

The goal of the task is a classification of text messages from users who can present some kind of cognitive distortion into one of the 10 predefined categories of cognitive distortions. It should be noted that this is a major challenge due to the overlap between the different classes of cognitive distortions defined in the psychological literature (Yurica and DiTomasso, 2005). Additionally, the complexity of natural language often leads to ambiguous cases, where messages could reasonably belong to more than one category. This leads to subjectivity playing an important role when features of different distortions can be noted in a message. For instance, Emotional Reasoning and Fortune-telling can both involve assumptions about future outcomes, while Overgeneralization and Labeling may both rely on broad or reductive statements. This also makes the inter-annotator agreement low and therefore, the result that can be obtained from an automatic classification is limited.

In this study, classification will be performed on the 10 classes of cognitive distortions present in the dataset:

- *Emotional reasoning:* formulating arguments based on feelings rather than objective reality.
- *All-or-nothing thinking (polarized thinking):* interpreting events and people in absolute terms such as “always”, “never”, or “everyone”, without justification.
- *Overgeneralization:* drawing broad conclusions from isolated cases and assuming their validity in all contexts.
- *Mind reading:* assuming others’ intentions or thoughts without empirical evidence.
- *Fortune-telling:* predicting future events with certainty, often emphasizing negative outcomes while ignoring available evidence.
- *Magnification:* exaggerating the worst possible outcomes or downplaying the severity of a situation, imagining it as unbearable when it is merely uncomfortable or inconvenient.
- *Should statements:* imposing rigid rules on oneself or others, often leading to feelings of frustration or inadequacy.
- *Labeling:* assigning a generalized and often negative label to oneself or others, typically using the verb “to be”.
- *Personalization:* assuming responsibility for events or believing that others’ actions are directly related to oneself, whether positively or negatively.
- *Mental filter:* obsessively focusing on a single negative aspect to the exclusion of all other qualities or circumstances.

A brief description of these can be found in the dataset repository or in their original article (Shreevastava and Foltz, 2021). In addition, the annotators of this dataset identified the part of the whole message that they considered the cognitive distortion to be manifested. For this study, we considered using the part containing the cognitive distortion of the messages.

Due to the overlap between classes, in the dataset used, annotators were asked to always label the distortion they found dominant in the text, but they also had the possibility to annotate a secondary cognitive distortion when they thought it relevant. Thus, the evaluation criteria followed in this study consider a prediction correct if it matches either the dominant or the secondary distortion.

### 3.1.2 Prompt Building

When using text-generating LLMs for specific tasks, the methodology used in constructing the prompt is very important. The prompt is the instructions to be followed by the model to perform the specific task and to structure the output in a certain way. This prompt and the text to be classified is the input received by the models. The performance of the model on the task will be greatly influenced by the prompt constructed and used, as it is what guides the model in constructing its response.

To guide the classification process, the prompt includes:

- Task Contextualization and Role Definition: The prompt begins by defining the role of the LLM as a Cognitive Distortion Classifier with the knowledge of an expert psychologist. A brief explanation of what cognitive distortions are and their relevance is provided to contextualize the task.
- A task description: The models are instructed to identify the most dominant distortion present in a hypothetical message from a dataset, emphasizing that only one category should be selected.
- A comprehensive list of cognitive distortion categories: Each category is defined with examples to illustrate its application. This serves as both a taxonomy and a guide for the models to differentiate between similar distortions. The definition of each cognitive distortion is the one given by the authors of the dataset in the repository, to use the same ones on which the annotation was based.
- Output specification: The desired response format is explicitly defined as a JSON object containing only the category of the identified distortion.

This structured approach was applied consistently across all LLMs evaluated, ensuring comparability in their outputs. The prompt also incorporates examples of messages for each cognitive distortion category, further clarifying distinctions and reducing ambiguity for the models. This design aimed to maximize the models' ability to recognize the subtle differences between overlapping categories, such as Overgeneralization and Mental Filter, or Fortune-telling and Emotional Reasoning.

### 3.1.3 Computational Resources

To evaluate text-generating LLMs for the classification of cognitive distortions, the experiments in this study were conducted using a high-performance computing setup. The hardware resources employed included an NVIDIA A40 GPU with 48GB VRAM, an AMD EPYC 7453 processor with 28 cores, and 512GB of RAM. This configuration ensured efficient processing and evaluation of the large-scale models tested in this study.

### 3.1.4 Dataset

In this study, we evaluate the capability of text-generating LLMs in detecting cognitive distortions within patient-therapist interactions. To achieve this, we use an annotated dataset derived from the publicly available Therapist Q&A dataset[1], which contains anonymized question-and-answer exchanges between patients and licensed therapists. Each patient input typically describes their situation, symptoms, or thoughts, which are then addressed by a therapist's response. The Therapist Q&A dataset was labeled with ten cognitive distortions as described in Section 3.1.1. This dataset comprises 2,530 annotated samples of patient inputs, each accompanied by a dominant distortion label, and in some cases, an optional secondary distortion. The labels correspond to ten common cognitive distortions identified in Cognitive Behavioral Therapy (CBT): All-or-nothing thinking, Overgeneralization, Mental filtering, Should statements, Labeling, Personalization, Magnification, Emotional reasoning, Mind reading, and Fortune-telling. If no cognitive distortion was detected in a sample, it was labeled as "No distortion". Annotators were also tasked with highlighting specific sentences in the patient inputs that indicated the presence of distorted reasoning, providing crucial context for interpretation.

### 3.1.5 Selected LLM Models

In this section, we present the LLMs chosen for our comparative study on detecting cognitive distortions in text user messages. The selection criteria prioritized relevance, backing by reputable organizations, and performance in tasks such as reasoning, text comprehension, and instruction adherence. To ensure a balanced evaluation, we included models of varying sizes, architectures, and intended use cases. Including diversity allows us to analyze how different configurations impact the models' ability to address the nuanced task of detecting cognitive distortions.

The models chosen in this work are:

- Google's Gemma 2 family (Riviere et al., 2024) includes compact models optimized for NLP and text generation tasks. For this study, we used the 2B and 9B parameter models. They have 8k token context window and are trained on diverse datasets, balancing computational efficiency with ethical safeguards to minimize risks and biases.
- Meta's Llama 3 series (Grattafiori et al., 2024) provided distinct subsets. From the Llama 3.1 family, we selected the 8B and 70B parameter models, designed for high-performance environments. The smaller 1B and 3B models from the Llama 3.2 family, optimized for constrained settings, were also included. Notably, Llama 3.2 leverages knowledge distillation from larger models, and both series support a 128k token context window, except in the quantized versions of Llama 3.2 models.
- The Mistral AI models (Jiang et al., 2023) included were the Ministral 8B and the Mistral NeMo (12B). The Ministral 8B is a compact, multilingual model optimized for low-latency tasks, while the NeMo (12B) excels in reasoning and supports over 100 languages. Both models feature a 128k token context window, offering flexibility for long-context tasks.
- From Microsoft's Phi-3 family (Abdin et al., 2024), we selected two models. The Phi-3.5 Mini (3.8B) supports a 128k token context and is well-suited for multilingual reasoning tasks. The Phi-3 Medium (14B), though limited to a 4k token context, performs exceptionally in logic and math-related benchmarks, particularly in English.
- Finally, we selected several models from Alibaba Cloud's Qwen 2.5 series (Yang et al., 2025), particularly the variants of parameters 1.5B, 3B, 7B and 14B. While the smaller models (1.5B and 3B) handle structured tasks within a 32k token context, the larger variants (7B and 14B) support up to 128k tokens. These models excel in generating structured outputs such as JSON, along with advanced reasoning.

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

# 4 RESULTS

This section presents the evaluation results obtained in our study for the selected models in the task of detecting cognitive distortions in users' messages. The results, detailed in Table 1 and Figure 1, show clear trends related to model families and parameter sizes.

The Gemma 2 models showed a predictable relationship between size and performance. The smaller Gemma-2-2B achieved modest results with 0.2 accuracy and an F1 score of 0.15. In contrast, the larger Gemma-2-9B performed significantly better, reaching 0.33 accuracy and 0.25 F1. This highlights the scalability of this model family.

In the Llama family, improvements were also tied to size. Llama-3.2-1B performed at the lower end, with 0.11 accuracy and 0.09 F1. However, Llama-3.2-3B, showed notable progress, achieving 0.23 and 0.19, respectively. Llama-3.1-8B matched Gemma-2-9B with 0.33 accuracy and 0.28 F1. At the top of the family, Llama-3.1-70B emerged as the overall leader achieved 0.39 accuracy and 0.35 F1, underscoring its capacity for complex tasks.

The Mistral series maintained solid and consistent performance across models. The Ministral-8B improved slightly on the Mistral-7B, achieving 0.29 accuracy and 0.25 F1. Meanwhile, the larger Mistral Nemo (12B) reached 0.37 accuracy and 0.31 F1, demonstrating its strength in handling nuanced tasks despite not being the largest model overall.

Microsoft's Phi-3 models delivered competitive results. The Phi-3.5 Mini (3B) stood out with 0.3 accuracy and 0.24 F1, surpassing some larger models in other families. Interestingly, the Phi-3 Medium (14B) achieved very similar results, with 0.32 accuracy and 0.27 F1 suggesting this time that differences in training or architecture may be an important factor in tasks such as the one proposed in this study.

The Qwen 2.5 models, like most, showed steady progress with size and competitive results in general. The smallest, Qwen-2.5-1.5B, achieved 0.17 accuracy and 0.1 F1. Larger models, such as the 3B and 7B variants, saw performance jump to 0.21/0.16 and 0.32/0.35, respectively. The largest model, Qwen-2.5-14B, joined the top tier with 0.4 accuracy and 0.42 F1, emerging as the overall leader. making it one of the most effective in the study.

Across all results, Llama-3.1-70B and Qwen-2.5-14B led the group. However, small-sized and mid-sized models like Phi-3.5 Mini (3B) and Mistral Nemo (12B) also deserve recognition for their competitive performance relative to their size and efficiency.

# 5 DISCUSSION

The results of the study have shown several interesting insights. Despite not being trained for the recognition of cognitive distortions, and for this reason achieving relatively modest performance, the results obtained by the LLMs in this study are promising. These results demonstrate the potential of LLMs to identify patterns indicative of cognitive distortions, even without task-specific optimization and suggests that, with appropriate fine-tuning, they could become valuable tools.

A first insight is that within the same family of models, where the models follow a similar structure and training, a clear trend is detected in which the results for the cognitive distortion recognition task improve with the increase in size of the models. For instance, within the Llama family, the 70B model achieved the highest results, significantly surpassing its smaller counterparts. Similarly, the Qwen-2.5-14B model delivered outstanding results, well above the rest of the Qwen models. This confirms that the better ability of the larger models to detect more subtle nuances in texts is linked to a better understanding of mental health issues and therefore, to the classification of cognitive distortions. On the other hand, it is also evident that differences in model architectures or training are also related to better model performance and better outcome on this task. This is well exemplified by the Phi3.5 mini model, which not only outperforms models in the same size range (3B parameters), but also approaches and even outperforms some of the medium-sized models in this experiment, as Ministral with 8B. A noteworthy feature of Phi3.5 is that its training is oriented toward reasoning, logic, and mathematics. This focus could have made it more effective in identifying implicit patterns and underlying relationships in texts. This could be prepared this model better to detect relevant nuances in texts with cognitive distortions, where subtle differences in language may be decisive. A similar pattern can be found in the Qwen 2.5 model series, with Qwen 2.5 14b outperforming Llama 70b, whose training emphasized knowledge, coding and mathematics. This suggests that efficient architectures and specialized training can enhance performance, offering alternatives to relying solely on increased parameter counts. Therefore, for applications such as the recognition of cognitive distortions, smaller but specialized model could offer superior performance compared to larger and more generalist models.

Another insight is that the smallest models, such as Llama-3.2-1B and Qwen-2.5-1.5B, had difficulty capturing the complexity of the task, probably due

Table 1: Performance of Models. The metrics are obtained through macro-averaging.

| **Model** | **Size** | **Accuracy** | **F1** | **Recall** | **Precision** |
|---|---|---|---|---|---|
| gemma-2-2b-it | 2b | 0.20 | 0.15 | 0.17 | 0.40 |
| gemma-2-9b-it | 9b | 0.33 | 0.25 | 0.29 | 0.28 |
| Llama-3.2-1B-instruct | 1b | 0.11 | 0.09 | 0.11 | 0.12 |
| Llama-3.2-3B-instruct | 3b | 0.23 | 0.19 | 0.20 | 0.27 |
| Llama-3.1-8B-instruct | 8b | 0.33 | 0.28 | 0.31 | 0.36 |
| Llama-3.1-70b-instruct | 70b | **0.39** | **0.35** | **0.37** | **0.41** |
| Mistral-7B-Instruct-v0.3 | 7b | 0.23 | 0.20 | 0.19 | 0.32 |
| Ministral-8B-Instruct-2410 | 8b | 0.29 | 0.25 | 0.28 | 0.45 |
| Mistral-Nemo-Instruct-2407 | 12b | 0.37 | 0.31 | 0.35 | 0.52 |
| Phi-3.5-mini-instruct | 3b | 0.30 | 0.24 | 0.27 | 0.33 |
| Phi-3-medium-4k-instruct | 14b | 0.32 | 0.27 | 0.28 | 0.33 |
| Qwen2.5-1.5B-Instruct | 1.5b | 0.17 | 0.10 | 0.15 | 0.23 |
| Qwen2.5-3B-Instruct | 3b | 0.21 | 0.16 | 0.19 | 0.20 |
| Qwen2.5-7B-Instruct | 7b | 0.32 | 0.35 | 0.35 | 0.38 |
| Qwen2.5-14B-Instruct | 14b | **0.40** | **0.42** | **0.43** | **0.50** |

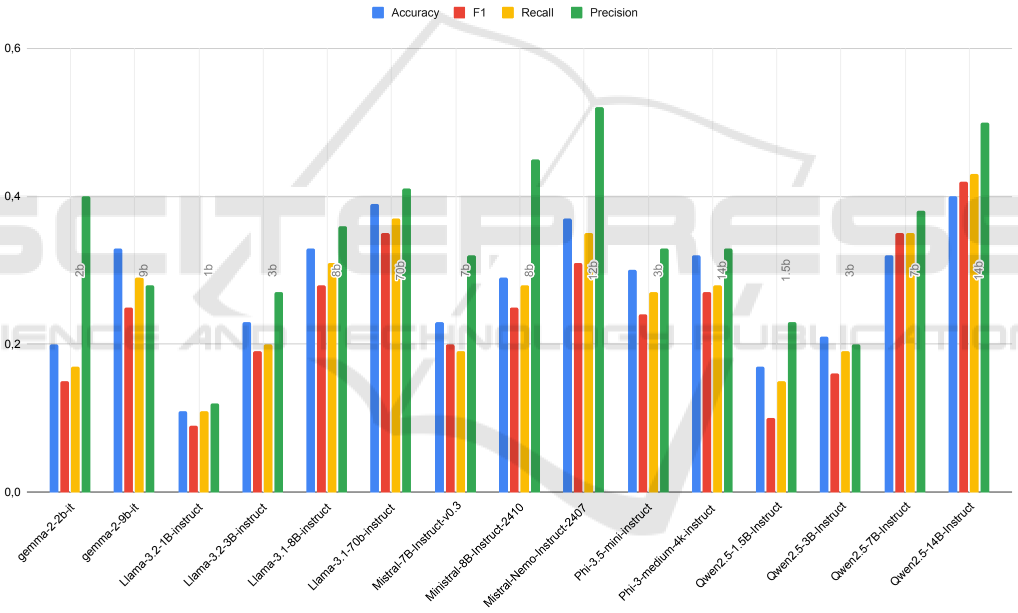

Figure 1: Performance of models on the cognitive distortions classification task. The results include accuracy, F1, recall and precision metrics, showing the impact of model size and family on task performance. The metrics are obtained through macro-averaging.

to their limited capacity and contextual understanding. Although these models are efficient, their performance highlights trade-offs between resource constraints and task-specific requirements. Therefore, to use these types of models in edge computing we need to find a balance between reduced size and speed versus the performance of the model on the task. This balance can come from approaches such as Phi3.5 mini, which has achieved remarkable results while being lightweight. This size of the models seems like a good option to tune them and help them achieve better results without compromising that they can run on the device.

A critical common limitation of the dataset for cognitive distortion recognition and task itself was confirmed by this study: the subjectivity inherent in the annotation of cognitive distortions. This is evidenced by the frequently low inter-annotator agreement (33.7% in the case of the dataset used), reflecting the overlap of cognitive distortion categories. For example, statements classified as “emotional reasoning” may also correspond to “fortune-telling” depending on interpretation. The ambiguous nature of the task poses significant challenges and limits the maximum achievable performance of an automatic classification with the models. This highlights the need to redefine the different categories or to improve the design of the datasets and the task itself. For example, providing clearer guidelines for annotation or exploring multi-label classification approaches could help mitigate this problem.

The results suggest several implications for practical applications. Larger models such as Llama-3.1-70B and Qwen-2.5-14B are well suited for high-resource environments, delivering state-of-the-art performance. However, efficient models such us Phi-3.5 Mini offer promising alternatives for resource-constrained settings, especially when paired with task-specific optimizations.

## 6 ETHICS

The integration of LLMs in clinical mental health applications, such as the proposed classification of cognitive distortions, presents major challenges in terms of ethics and ensuring patient safety and privacy. On the one hand, too much reliance on these systems carries the risk of misdiagnosis or therapeutic recommendations due to possible misclassification of the models. These could appear due to imbalances in the training data, which could cause these systems to be affected by biases. Therefore, these systems should be a complementary aid for professionals and not a substitute for their clinical expertise. Another risk is in the training of the models, as they may memorize personal and sensitive patient data from the conversations used as examples for the models. Therefore, strict privacy protections, such as data anonymization and on-device processing, must be applied.

## 7 CONCLUSIONS AND FUTURE WORK

In this paper we have proposed the use of text-generating LLMs as classifiers of cognitive distortions, because of their potential to analyze and extract details and patterns from texts. Although these models originated on a large scale and needing a great computational power, smaller and more efficient models are being developed, which give us the opportunity to use their benefits in the analysis and generation of natural language, as well as their great versatility to adapt to different tasks, in environments with more limited resources or even on device.

In our experiments, larger models demonstrated generally superior performance, confirming that increased model size enhances the ability to capture nuanced text features relevant to mental health tasks. However, the results also provide evidence that smaller models may have an interesting trade-off between efficiency and performance. In addition, some models have been shown to have similar or better performance for the task than larger models, which suggests that there are other very influential factors such as the training focus.

In future work, fine-tuning an LLM specifically for the task of detecting cognitive distortions presents a promising way for improvement, as the models in this study were not trained for this specific task, but were based on their general-purpose capabilities. Fine-tuning could significantly improve the performance of the models by better adjusting them to the nuances of this domain. This approach is likely to be especially beneficial with small models, such as Phi-3.5 Mini or Llama-3.2-3B, which demonstrated competitive performance despite their size, as we could obtain competitively performing classifiers that are lightweight and fast. In addition, improving the quality of the dataset remains a key area of future work. A more precise definition of the different categories of cognitive distortions or the adoption of multi-label classification frameworks could solve the problem of label overlap and ambiguity.

## ACKNOWLEDGEMENTS

Work was partially supported by Generalitat Valenciana CI-PROM/2021/077, Spanish government PID2023-151536OB-I00 and First Research Project Support (PAID-06-24) by Vicerrectorado de Investigación de la Universitat Politècnica de València.

## REFERENCES

Abdin, M., Aneja, J., Awadalla, H., Awadallah, A., Awan, A. A., Bach, N., Bahree, A., Bakhtiari, A., Bao, J., Behl, H., et al. (2024). Phi-3 technical report: A highly capable language model locally on your phone. 10.48550/arXiv.2404.14219.

Alhaj, F., Al-Haj, A., Sharieh, A., and Jabri, R. (2022). Improving arabic cognitive distortion classification in twitter using bertopic. *International Journal of Advanced Computer Science and Applications (IJACSA)*, 13(1).

Annepaka, Y. and Pakray, P. (2024). Large language models: a survey of their development, capabilities, and applications. *Knowledge and Information Systems*, TBD(TBD):TBD.

APA (2024). American Psychological Association Dictionary of Psychology. https://dictionary.apa.org/cognitive-distortion. Last accessed 2024-12-01.

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

Beck, A. T. (1964). Thinking and depression: Ii. theory and therapy. *Archives of General Psychiatry*, 10(6):561–571.

Beck, J. S. (1995). *Cognitive therapy: Basics and beyond*. Guilford Press.

Blake, E., Dobson, K. S., Sheptycki, A. R., and Drapeau, M. (2016). The relationship between depression severity and cognitive errors. *American Journal of Psychotherapy*, 70(2):203–221. PMID: 27329407.

Day, A. (2017). Cognitive-behavioural therapy. *Individual Psychological Therapies in Forensic Settings*, pages 28–40.

Ding, X., Lybarger, K., Tauscher, J. S., and Cohen, T. (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.

Fortune, E. E. and Goodie, A. S. (2012). Cognitive distortions as a component and treatment focus of pathological gambling: a review. *Psychology of addictive behaviors*, 26(2):298.

Grattafiori, A., Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Yang, A., et al. (2024). The llama 3 herd of models. 10.48550/arXiv.2407.21783.

Hossain, S. B. (2009). *Understanding Patterns of cognitive Distortions*. PhD thesis, M. Phil Thesis submitted to the Dept. of Clinical Psychology, DU.

Jager-Hyman, S., Cunningham, A., Wenzel, A., Mattei, S., Brown, G. K., and Beck, A. T. (2014). Cognitive distortions and suicide attempts. *Cognitive Therapy and Research*, 38(4):369–374.

Jiang, A. Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D. S., de las Casas, D., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., Lavaud, L. R., Lachaux, M.-A., Stock, P., Scao, T. L., Lavril, T., Wang, T., Lacroix, T., and Sayed, W. E. (2023). Mistral 7b. 10.48550/arXiv.2310.06825.

Kira, K. and Rendell, L. A. (1992). A practical approach to feature selection. In Sleeman, D. and Edwards, P., editors, *Machine Learning Proceedings 1992*, pages 249–256. Morgan Kaufmann, San Francisco (CA).

Kononenko, I. (1994). Estimating attributes: Analysis and extensions of relief. In Bergadano, F. and De Raedt, L., editors, *Machine Learning: ECML-94*, pages 171–182, Berlin, Heidelberg. Springer Berlin Heidelberg.

Mahali, S. C., Beshai, S., Feeney, J. R., and Mishra, S. (2020). Associations of negative cognitions, emotional regulation, and depression symptoms across four continents: International support for the cognitive model of depression. *BMC Psychiatry*, 20(1):18.

Mostafa, M., El Bolock, A., and Abdennadher, S. (2021). Automatic detection and classification of cognitive distortions in journaling text. In *Proceedings of the 17th International Conference on Web Information Systems and Technologies - WEBIST*, pages 444–452. SciTePress.

Riviere, M., Pathak, S., Sessa, P. G., Hardin, C., Bhupatiraju, S., Hussenot, L., Mesnard, T., Shahriari, B., Ramé, A., et al. (2024). Gemma 2: Improving open language models at a practical size. https://doi.org/10.48550/arXiv.2408.00118.

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

Shreevastava, S. and Foltz, P. (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.

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

Tauscher, J. S., Lybarger, K., Ding, X., Chander, A., Hudenko, W. J., Cohen, T., and Ben-Zeev, D. (2023). Automated detection of cognitive distortions in text exchanges between clinicians and people with serious mental illness. *Psychiatric services*, 74(4):407–410.

Uban, A.-S., Chulvi, B., and Rosso, P. (2021). An emotion and cognitive based analysis of mental health disorders from social media data. *Future Generation Computer Systems*, 124:480–494.

Wang, B., Deng, P., Zhao, Y., and Qin, B. (2023). C2d2 dataset: A resource for analyzing cognitive distortions and its impact on mental health. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10149–10160.

Yang, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Li, C., Liu, D., Huang, F., ..., H. W., and Qiu, Z. (2025). Qwen2.5 technical report. https://doi.org/10.48550/arXiv.2412.15115.

Yurica, C. L. and DiTomasso, R. A. (2005). Cognitive distortions. *Encyclopedia of cognitive behavior therapy*, pages 117–122.