ERD: A Framework for Improving LLM Reasoning for Cognitive Distortion Classification

Sehee Lim[*]
Yonsei University
sehee0706@yonsei.ac.kr

Yejin Kim[*]
Yonsei University
yjkim.stat@yonsei.ac.kr

Chi-Hyun Choi[*]
EverEx
leo@everex.co.kr

Jy-yong Sohn[†]
Yonsei University
EverEx
jysohn1108@yonsei.ac.kr

Byung-Hoon Kim[†]
Yonsei University
EverEx
egyptdj@yonsei.ac.kr

Abstract

Improving the accessibility of psychotherapy with the aid of Large Language Models (LLMs) is garnering a significant attention in recent years. Recognizing cognitive distortions from the interviewee's utterances can be an essential part of psychotherapy, especially for cognitive behavioral therapy. In this paper, we propose ERD, which improves LLM-based cognitive distortion classification performance with the aid of additional modules of (1) extracting the parts related to cognitive distortion, and (2) debating the reasoning steps by multiple agents. Our experimental results on a public dataset show that ERD improves the multi-class F1 score as well as binary specificity score. Regarding the latter score, it turns out that our method is effective in debiasing the baseline method which has high false positive rate, especially when the summary of multi-agent debate is provided to LLMs.

1 Introduction

Large Language Models (LLMs) are dominating the research areas in machine learning and artificial intelligence, broadening its usage in various applications (Radford et al., 2018, 2019; Brown et al., 2020; OpenAI, 2023; Ouyang et al., 2022). Especially in the medical domain, PaLM (Chowdhery et al., 2022) and its variants, such as Med-PaLM (Singhal et al., 2022), are equipped with medical data and instructions to answer the questions from clinical field (Chowdhery et al., 2022; Singhal et al., 2023). In addition, conversational AI assistant chatbots are devised to support patients with mental health issues (Rathje et al., 2023; Vaidyam et al., 2019; Saha et al., 2022; Stock et al., 2023; Liu et al., 2023; Welivita et al., 2021; Sharma et al., 2020).

Recognizing the fact that individuals with mental disorders hesitate to seek in-person medical consultations (Steinberg et al., 1980), previous studies (Yang et al., 2023; Lee et al., 2023; Chen et al., 2023b) attempt to enhance the accessibility and quality of psychotherapy through the use of LLMs with Chain-of-Thought (CoT) reasoning (Wei et al., 2022). These models aim to detect the user's personality and interpret their mental state in order to generate more empathetic responses.

For example, Diagnosis-of-Thought (DoT) uses LLMs to classify cognitive distortions from utterances, which is a crucial part of Cognitive Behavior Therapy (CBT) (Chen et al., 2023b).

While the DoT method holds promise, one key challenge that remains an open issue is the tendency of the model to overdiagnose cognitive distortions, incorrectly inferring irrational thought patterns even when the user's statements are benign. In addition, the distortion classification performance of DoT in multi-class setup is close to that of random guessing, which limits its usage in practice.

In this paper, we tackle these issues by proposing a new framework for classifying cognitive distortions from the user utterances, by introducing modules for debiasing the overdiagnosing tendency of existing methods and for improving the performance on classifying distortion types inferred from the utterances.

Our main contributions can be summarized as below:

- We introduce ERD, a new framework for classifying cognitive distortions in the user utterances using three steps: Extraction, Reasoning, and Debate, each of which uses LLMs. The first step lets LLM extract a part of the utterances that is related with the distortion, the second step uses LLM to generate the thought process of estimating cognitive distortions from the extracted part, and the third step uses multi-agent LLMs to discuss the thought process described in the second

[*]Equal contribution
[†]Corresponding authors

Proceedings of the 6th Clinical Natural Language Processing Workshop, pages 292–300

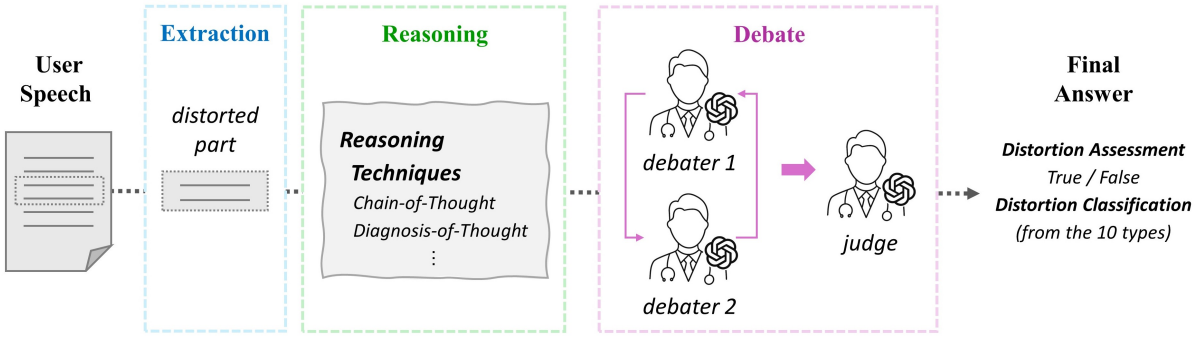

Figure 1: The pipeline of Extraction-Reasoning-Debate (ERD), which detects and classify the cognitive distortion from the input user speech. It begins with the identification and extraction of potential cognitive distortions from the user speech. These extracted elements are then utilized to construct an intermediate reasoning step. Subsequently, a debate is conducted, wherein multiple LLM agents deliberate to assess the presence and type of cognitive distortion. Finally, a judge integrates the entire debate process to get the final answer on the distortion classification problem.

step and make the final decision.

- Compared with existing baselines, ERD improves the multi-class F1 score for distortion classification task by more than 9% and improves the distortion assessment specificity score by more than 25%, when tested on the cognitive distortion detection dataset with 2530 samples in Kaggle.

- We provide factor analysis on ERD, showing that (1) multiple rounds of debate in ERD is beneficial for improving the classification score, and (2) the summarization and the validity evaluation processes during the debate step enhance the debiasing effect.

2 ERD

We propose Extraction-Reasoning-Debate (ERD), a framework for classifying distortions in a given user speech, as shown in Fig. 1. The prompts we used can be found in Figure 3 in Appendix. Below we elaborate each step in our framework.

Input	Distortion Classification
User Speech	$15.28_{0.65}$
Distorted Part of User Speech	$\mathbf{27.08}_{0.27}$

Table 1: Multi-class F1 score of DoT (Chen et al., 2023b) for the cognitive distortion classification problem, when two different inputs are given. The first option uses the user speech as the input, as done in (Chen et al., 2023b). The second option is considered by us, which only puts the ground-truth distorted part within the user speech. Putting only the distorted part significantly improves the classification performance, which motivates the Extraction step in ERD framework.

2.1 Extraction

To provide the motivation for the Extraction step proposed in our method, we first share our empirical results showing that extracting the distorted parts of user speech is beneficial for distortion classification. Table 1 shows the multi-class F1 score of Diagnosis-of-Thought (DoT) method for distortion classification problem (predicting out of 10 classes), tested on a cognitive distortion detection dataset with 2530 samples in Kaggle[1]. We test on two different options: (1) putting the user speech as it is, and (2) putting the ground-truth part (provided in the 'distorted part' column of the dataset) within the speech, that indicates the distortion. Table 1 shows that the multi-class F1 score increases more than 10% when the ground-truth distorted part is extracted before running DoT.

Motivated by this result, prior to the Reasoning step (e.g., DoT) which outputs the thought process for assessing/classifying the distortion, we add an Extraction step which instructs LLMs isolate the segments from the user's utterance that may potentially exhibit cognitive distortions. This process of extraction is done *without* paraphrasing or summarizing, thereby preserving the original context and nuances for the subsequent thought process. In summary, Extraction process ensures that the LLMs' responses hinge on the most informative facets of the utterance, which in turn enhance the quality of the distortion classification performance.

2.2 Reasoning

Our target task (cognitive distortion classification from the user speech) is naturally considered as a

[1] https://www.kaggle.com/datasets/sagarikashreevastava/cognitive-distortion-detetction-dataset

Method	Distortion Assessment (True/False) Sensitivity	Specificity	F1 Score	Distortion Classification (out of 10 types) Weighted F1 Score
Reasoning	$\underline{99.29}_{0.19}$	$6.79_{0.34}$	$\mathbf{78.26}_{0.16}$	$15.28_{0.65}$
+Extraction	$\mathbf{99.83}_{0.03}$	$0.93_{0.22}$	$\underline{77.48}_{0.04}$	$\mathbf{24.40}_{0.69}$
+Debate	$73.10_{0.26}$	$\mathbf{33.05}_{0.58}$	$68.89_{0.24}$	$22.18_{0.99}$
+Extraction+Debate	$74.89_{2.31}$	$\underline{30.74}_{3.92}$	$69.49_{0.62}$	$\underline{24.27}_{1.14}$

Table 2: Cognitive distortion assessment/classification results of ERD when various modules (Extraction and Debate) are added. Here, we test on cognitive distortion detection dataset in Kaggle, and use DoT (Chen et al., 2023b) method for the Reasoning step. Upon the above results, Extraction improves the distortion classification performance and Debate increases the distortion assessment specificity significantly. Combining both Extraction and Debate takes the sweet spot, simultaneously enhancing both performances.

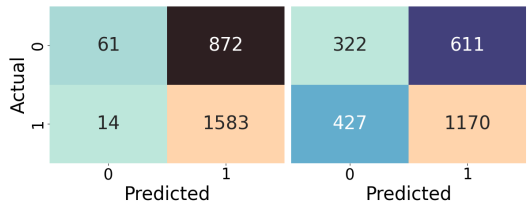

Figure 2: Confusion matrices of ERD when tested on 2530 samples: (Left) only Reasoning is used, (Right) Extraction, Reasoning and Debate steps are used. Including Extraction and Debate modules increases the number of true negatives from 61 to 322, thus correctly identifying the samples with 'no distortion'.

task that requires logical thinking, if we imagine how doctors classify the patients. In recent years, various methods propose letting LLMs mimic the logical thought process or reasoning steps. For example, chain-of-thought (CoT) prompting and its variants (Wei et al., 2022; Kojima et al., 2022; Yao et al., 2023; Besta et al., 2023; Chen et al., 2023a; Yang et al., 2023; Lee et al., 2023) provide a significant performance improvement in various reasoning tasks including common sense reasoning and mathematical reasoning.

Our Reasoning step chooses any existing methods which let LLMs output the thought process for performing the target task. By default, we use diagnosis-of-thought (DoT) (Chen et al., 2023b) comprised of three critical stages (subjectivity assessment, contrastive reasoning, and schema analysis) that construct rationales for the detection of cognitive distortions. At the *subjectivity assessment* stage, the input utterances are differentiated between the objective facts and the subjective thoughts. This is followed by the *contrastive reasoning* stage, where the process elicits both supportive and contradictory perspectives to the speaker's viewpoint. The final stage, *schema analysis*, involves delving into the underlying thought schema, which refers to the subconscious cognitive patterns or frameworks that shape and influence a person's specific thought process and behavior.

2.3 Debate

Several recent works on using LLMs for reasoning tasks show that multiple LLM agents debating their thought processes significantly improve the performance (Liang et al., 2023; Zheng et al., 2023; Xiong et al., 2023; Chan et al., 2023; Du et al., 2023). Motivated by this observation, we add multi-agent debate (or Debate) step following the Reasoning step. In Figure 1, ERD employs three LLM agents, each designated with the role of "physician" to simulate a professional medical debate. The discussion between first two agents (two debators) is overseen by the third agent (called the judge agent), bearing the role of "head doctor" who monitors the entire debate to ensure a fair evaluation. The third agent is introduced, motivated by recent result showing that LLMs can behave as a good judge (Zheng et al., 2023). The first debater presents arguments for the presence or absence of cognitive distortion in the user speech, based on the LLM outputs obtained in the Extraction and Reasoning steps. Subsequently, the second debator counters the initial assertions, presenting a contradicting viewpoint. The first debater then responds to this counterargument, followed by a second round of rebuttal from the second debater, resulting in two rounds of argumentation. One can consider repeating this iterative exchange of thoughts for multiple rounds. After this iterative process, the judge agent integrates the entire discourse, employing two proposed methodologies to reach a final decision.

We consider two different options for controlling the behavior of the judge agent to get better performances. The first option involves a straightforward summarization of the total debate process. The second option involves summarizing the debate and evaluating which side's arguments are more valid. By adding such summarization process, we expect that the final answer of ERD is based on a comprehensive consideration of all presented viewpoints.

Distortion Type	Count
All-or-nothing thinking	100
Emotional Reasoning	134
Fortune-telling	143
Labeling	165
Magnification	195
Mental filter	122
Mind Reading	239
Overgeneralization	239
Personalization	153
Should statements	107
No Distortion	933
In Total	2530

Table 3: Details of dataset used in this paper; we report the number of samples for each class including 10 types of distortions and "no distortion" type whose utterance does not contain distortion.

3 Experiments

Settings We use a cognitive distortion detection dataset (Shreevastava and Foltz, 2021) composed of speeches that correspond to 10 types of "cognitive distortions" and neutral speeches categorized as "no distortion" type. This dataset, sourced from Kaggle[1], contains 2530 annotated examples by experts and the least number of examples for each type of distortion is 100, which can be found in Table 3. This dataset is designed to facilitate two tasks: distortion assessment and distortion classification.

In the distortion assessment task, the model determines whether cognitive distortion is present in the patient's utterance. In the distortion classification task, the model identifies the specific type of cognitive distortion. We report the Sensitivity, Specificity and F1 score for the distortion assessment task, and the weighted F1 score for the distortion classification task. We run 3 random trials and report the mean and standard deviation values. We employ the model gpt-3.5-turbo with the temperature as 0.1. Every result reported in this paper is based on the zero-shot prompting.

Experimental results Table 2 shows the performances of ERD, when different modules are plugged in. Compared with naive method using Reasoning module only, adding Extraction module improves the distortion classification score by more than 9%, and adding Debate module not only improves the distortion classification score by around 7%, but also improves the distortion assessment specificity by more than 25%. Fig. 2 shows the confusion matrix of the ERD for two cases: (1) when only the Reasoning module is used, and (2) when Extraction, Reasoning and Debate are used. This result shows that adding Extraction and Debate modules promotes the correct estimation of utterances with no distortion. This qualitative result can be supported by our qualitative results (in Fig. 4 and Fig.5 in Appendix) showing the effectiveness of Debate step for improving the estimation performance. For a given speech (that does not have cognitive distortion), Fig. 4 and Fig. 5 show the responses of LLM, when Debate step is in-activated and activated, respectively. While LLM without Debate incorrectly estimates that the speech contains cognitive distortion (of type "Labeling"), LLM with Debate correctly estimates that the speech does not contain cognitive distortions.

Recall that in Debate step of ERD, we consider different prompting techniques to control the behavior of the judge agent when making the final decision. Table 4 shows the effect of such prompts for three variants:

(1) "ERD without summarization" does not instruct judge to summarize the claims of debate and just directly make decision, (2) "ERD with summarization" instructs judge to summarize the claims before making the decision, and "ERD with summarization and validity evaluation" instructs judge to summarize and evaluate the claims of debate before making the decision. Note that the specificity is keep improved as we provide more detailed instructions to the judge agent.

Table 5 shows how the performance improves as we increase r, the number of Debate rounds used in ERD. The results show that increasing the number of Debate rounds led to enhancements in both the binary F1 score and the multi-class F1 score. The performance saturates after $r = 2$, thus better to use two rounds of debate considering the token efficiency. This finding aligns with the results presented in a related work on multi-agent debate of LLMs, demonstrating a similar pattern in the impact of the number of debate rounds on the model performance (Du et al., 2023).

4 Conclusion

We introduce ERD, a framework using LLMs to estimate the cognitive distortion contained in the user utterances through three steps: Extracting distorted parts within the utterances, Reasoning the estimation of the corresponding distortion classes, and Debating the initial estimation using multiple agents. Compared with existing baselines only having the reasoning step, including the extraction

	Distortion Assessment			Distortion Classification
	Sensitivity	Specificity	F1 Score	Weighted F1 Score
ERD without Summarization	**92.13**$_{0.38}$	11.01$_{0.66}$	**75.48**$_{0.23}$	**25.28**$_{0.46}$
+Summarization	86.10$_{0.58}$	19.58$_{1.36}$	73.88$_{0.36}$	23.96$_{1.05}$
+Summarization+Validity Evaluation	74.89$_{2.31}$	**30.74**$_{3.92}$	69.49$_{0.62}$	24.27$_{1.14}$

Table 4: Comparison of ERD with three different prompting options that control the behavior of the judge. For all three options, the Extraction and Reasoning modules are active in all cases, with differences applied exclusively to the Debate module. For the first option, judge predicts the cognitive distortion type only based on the debate process log, without any summarization step. For the second option, judge first *summarizes* the debate and then predicts the cognitive distortion type. In the final option, judge *summarizes* the debate, *evaluates the validity* of the claims in the debate, and then predicts the cognitive distortion type. Both summarization and validity evaluation steps improve the performance in terms of specificity. Note that the number of Debate rounds is set to $r = 2$.

Metric	Round 1	Round 2	Round 3
Binary F1	52.13$_{1.25}$	69.49$_{0.62}$	**70.74**$_{0.44}$
Multi-class F1	22.79$_{1.62}$	24.27$_{1.14}$	**24.83**$_{0.81}$

Table 5: F1 scores for different r, the number of Debate rounds. The performances improve as r increases.

and debating steps improve the distortion classification performance by 9% and improve the distortion assessment specificity by over 25%. Such improvements is crucial to cognitive behavior therapy since ERD is more adept at correctly identifying cases without distortions, avoiding the pitfall of over-diagnosing cognitive distortions. Furthermore, experimental results reveal that we can control the behavior of ERD with various prompting options.

References

Maciej Besta, Nils Blach, Ales Kubicek, Robert Gerstenberger, Lukas Gianinazzi, Joanna Gajda, Tomasz Lehmann, Michal Podstawski, Hubert Niewiadomski, Piotr Nyczyk, et al. 2023. Graph of thoughts: Solving elaborate problems with large language models. *arXiv preprint arXiv:2308.09687*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Chi-Min Chan, Weize Chen, Yusheng Su, Jianxuan Yu, Wei Xue, Shanghang Zhang, Jie Fu, and Zhiyuan Liu. 2023. Chateval: Towards better llm-based evaluators through multi-agent debate.

Wenhu Chen, Xueguang Ma, Xinyi Wang, and William W. Cohen. 2023a. Program of thoughts prompting: Disentangling computation from reasoning for numerical reasoning tasks.

Zhiyu Chen, Yujie Lu, and William Yang Wang. 2023b. Empowering psychotherapy with large language models: Cognitive distortion detection through diagnosis of thought prompting. *arXiv preprint arXiv:2310.07146*.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2022. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, and Igor Mordatch. 2023. Improving factuality and reasoning in language models through multiagent debate.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213.

Yoon Kyung Lee, Inju Lee, Minjung Shin, Seoyeon Bae, and Sowon Hahn. 2023. Chain of empathy: Enhancing empathetic response of large language models based on psychotherapy models. *arXiv preprint arXiv:2311.04915*.

Tian Liang, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, Zhaopeng Tu, and Shuming Shi. 2023. Encouraging divergent thinking in large language models through multi-agent debate.

June M Liu, Donghao Li, He Cao, Tianhe Ren, Zeyi Liao, and Jiamin Wu. 2023. Chatcounselor: A large language models for mental health support. *arXiv preprint arXiv:2309.15461*.

R OpenAI. 2023. Gpt-4 technical report. *arXiv*, pages 2303–08774.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. 2018. Improving language understanding by generative pre-training.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Steve Rathje, Dan-Mircea Mirea, Ilia Sucholutsky, Raja Marjieh, Claire Robertson, and Jay J Van Bavel. 2023. Gpt is an effective tool for multilingual psychological text analysis.

Tulika Saha, Saichethan Reddy, Anindya Das, Sriparna Saha, and Pushpak Bhattacharyya. 2022. A shoulder to cry on: Towards a motivational virtual assistant for assuaging mental agony. In *Proceedings of the 2022 conference of the North American chapter of the association for computational linguistics: Human language technologies*, pages 2436–2449.

Ashish Sharma, Adam S. Miner, David C. Atkins, and Tim Althoff. 2020. A computational approach to understanding empathy expressed in text-based mental health support.

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.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. 2022. Large language models encode clinical knowledge. *arXiv preprint arXiv:2212.13138*.

Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Le Hou, Kevin Clark, Stephen Pfohl, Heather Cole-Lewis, Darlene Neal, et al. 2023. Towards expert-level medical question answering with large language models. *arXiv preprint arXiv:2305.09617*.

Herbert Steinberg, Moshe Torem, and Stephen M Saravay. 1980. An analysis of physician resistance to psychiatric consultations. *Archives of General Psychiatry*, 37(9):1007–1012.

Anna Stock, Stephan Schlögl, and Aleksander Groth. 2023. Tell me, what are you most afraid of? exploring the effects of agent representation on information disclosure in human-chatbot interaction. In *International Conference on Human-Computer Interaction*, pages 179–191. Springer.

Aditya Nrusimha Vaidyam, Hannah Wisniewski, John David Halamka, Matcheri S Kashavan, and John Blake Torous. 2019. Chatbots and conversational agents in mental health: a review of the psychiatric landscape. *The Canadian Journal of Psychiatry*, 64(7):456–464.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837.

Anuradha Welivita, Yubo Xie, and Pearl Pu. 2021. A large-scale dataset for empathetic response generation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1251–1264.

Kai Xiong, Xiao Ding, Yixin Cao, Ting Liu, and Bing Qin. 2023. Examining inter-consistency of large language models collaboration: An in-depth analysis via debate.

Tao Yang, Tianyuan Shi, Fanqi Wan, Xiaojun Quan, Qifan Wang, Bingzhe Wu, and Jiaxiang Wu. 2023. Psycot: Psychological questionnaire as powerful chain-of-thought for personality detection. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L Griffiths, Yuan Cao, and Karthik Narasimhan. 2023. Tree of thoughts: Deliberate problem solving with large language models. *arXiv preprint arXiv:2305.10601*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena.

Debater 1's Instruction

You are a physician participating in a debate discussing the presence or absence of cognitive distortions in a given speech.
Your task is to 1) finish a few test of thoughts questions to analyze the thought patterns of the patient.
Then based on the test of thoughts analysis, 2) make an initial claim and 3) defend your position against another debater's refutation.

Debater 2's Instruction

You are a physician participating in a debate discussing the presence or absence of cognitive distortions in a given speech. Your task is to refute another debater's claim.
The other debater analyzed the person's speech and determined whether cognitive distortions were present, based on the analysis. They also identified specific types of cognitive distortions, if any were found.
Now it's your turn to refute the previous claim.

Judge's Instruction

You are a head doctor and your task is to monitor the entire debate and make a final decision.
In the debate, the two physicians discussed whether there was a cognitive distortion present and, if so, what type it was.

Debater 1) Reasoning

Given a speech of a person: "__*{person's speech}*__", finish the following diagnosis of thought questions:
0. Extract the sentence that most clearly shows distortion from the patient's utterance.
1. Based on the extracted sentence, answer the following question: what is the situation? Find out the facts that are objective; What is the person thinking or imagining? Find out the thoughts or opinions that are subjective.
2. What makes the person think the thought is true or is not true? Find out the reasoning processes that support and do not support these thoughts.
3. Why does the person come up with such reasoning process supporting the thought? What's the underlying cognitive mode of it?
Each stage should be constructed based on the findings of the preceding one.

Debater 1) Initial Claim

Then, based on the above analysis, Respond to the task.
1. Identify if there is cognitive distortion in the speech.
2. Recognize the specific types of the cognitive distortion.

Debater 2) Refutation

The opposing team's argument: "__*{debater1's initial claim}*__"
Refute the opposing team's arguments and present your new claim.

Debater 1) Defense

This is the opposing team's refutation of your opinion: "__*{debater2's refutation}*__"
Reaffirm and defend your original argument.

Debater 2) Second Refutation

This is the opposing physician's defense against your refute: "__*{debater1's defense}*__"
Counter this defense.

Judge) Summary and Reasoning

The discussion on whether there is cognitive distortion in this speech is as follows : "__*{entire debate process}*__"
Summarize the entire debate and evaluate which side presents a more valid argument.

Judge) Final Decision

Make a final decision.

Figure 3: Prompts of ERD proposed in this paper.

Speech

My mother and I have had a rocky history, mostly as a result of a childhood divorce and her alcohol addiction. My mother cheated on my father, who I always been extremely close to, and re-married immediately to a younger man. Subsequently over the years, she received treatment and changed her life, and so we re-established a relationship and have grown close.

Reasoning and Answering

1) There is a presence of cognitive distortion in the speech.
2) The specific cognitive distortion is "Labeling," as the patient reduces their mother to a single characteristic of betrayal and abandonment based on her actions. This oversimplification of the mother's behavior contributes to a distorted view of their relationship dynamics.

Figure 4: The output of `ERD` for a given speech, when the `Debate` step is not activated. LLM estimates that the speech contains cognitive distortion (type "Labeling") even though it does not.

Speech

My mother and I have had a rocky history, mostly as a result of a childhood divorce and her alcohol addiction. My mother cheated on my father, who I always been extremely close to, and re-married immediately to a younger man. Subsequently over the years, she received treatment and changed her life, and so we re-established a relationship and have grown close.

Debater1's Initial Claim

1) There is a presence of cognitive distortion in the speech.
2) The specific cognitive distortion is "Labeling," as the patient reduces their mother to a single characteristic of betrayal and abandonment based on her actions. This oversimplification of the mother's behavior contributes to a distorted view of their relationship dynamics.

Debater2's Refutation

I respectfully disagree with the opposing team's assessment. The patient's speech does not exhibit the cognitive distortion of labeling. The patient accurately describes specific actions of betrayal and abandonment by the mother, which are factual events, not an oversimplification or reduction of the mother to a single characteristic. The patient acknowledges the complexity of the relationship dynamics by also mentioning the mother's alcohol addiction and subsequent treatment, indicating a nuanced understanding of the situation. Therefore, there is no evidence of cognitive distortion in the form of labeling in the speech.

Debater1's Defense

I maintain that the patient's speech demonstrates the cognitive distortion of labeling. By focusing solely on the mother's actions of betrayal and abandonment, the patient reduces her to a single characteristic, overlooking the complexity of human behavior. While the events described may be factual, the patient's emphasis on these aspects distorts the overall perception of the relationship. The patient's strong emotional response to these specific actions indicates a biased interpretation, aligning with the cognitive distortion of labeling. Therefore, the presence of labeling in the speech is evident despite the acknowledgment of other factors.

Debater2's Second Refutation

I respectfully maintain my position that the patient's speech does not exhibit the cognitive distortion of labeling. The patient's detailed account of specific events does not equate to reducing the mother to a single characteristic. Acknowledging the complexity of the relationship dynamics by mentioning the mother's alcohol addiction and treatment shows a nuanced understanding, not oversimplification. The patient's emotional response to betrayal and abandonment does not automatically indicate labeling; it reflects a genuine reaction to significant events. Therefore, the presence of labeling in the speech remains unsubstantiated.

Judge's Summary and Reasoning

The debate revolves around whether the patient's speech exhibits the cognitive distortion of labeling. Debater 1 argues that the patient reduces the mother to a single characteristic of betrayal and abandonment, indicating labeling. On the other hand, Debater 2 contends that the patient's detailed account of events does not equate to labeling, as the patient acknowledges the complexity of the situation. While both sides make valid points, Debater 2's argument that the patient's nuanced understanding and emotional response do not necessarily indicate labeling appears stronger, as it considers the broader context and complexity of human behavior.

Judge's Final Decision

Presence of Cognitive Distortion: False
Classification of Cognitive Distortion Types: N/A

Figure 5: The output of `ERD` for a given speech (same as that shown in Fig. 4), when the `Debate` step is activated. LLM correctly estimates that the speech does not have cognitive distortions. This shows the debiasing effect of adding `Debate` step in `ERD`.