RU at WASSA 2024 Shared Task: Task-Aligned Prompt for Predicting Empathy and Distress

Haein Kong
Rutgers University
haein.kong@rutgers.edu

Seonghyeon Moon
Brookhaven National Laboratory
smoon@bnl.gov

Abstract

This paper describes our approach for the WASSA 2024 Shared Task on Empathy Detection and Emotion Classification and Personality Detection in Interactions at ACL 2024. We focused on Track 3: Empathy Prediction (EMP) which aims to predict the empathy and distress of writers based on their essays. Recently, LLMs have been used to detect the psychological status of the writers based on the texts. Previous studies showed that the performance of LLMs can be improved by designing prompts properly. While diverse approaches have been made, we focus on the fact that LLMs can have different nuances for psychological constructs such as empathy or distress to the specific task. In addition, people can express their empathy or distress differently according to the context. Thus, we tried to enhance the prediction performance of LLMs by proposing a new prompting strategy: Task-Aligned Prompt (TAP). This prompt consists of aligned definitions for empathy and distress to the original paper and the contextual information about the dataset. Our proposed prompt was tested using ChatGPT and GPT4o with zero-shot and few-shot settings and the performance was compared to the plain prompts. The results showed that the TAP-ChatGPT-zero-shot achieved the highest average Pearson correlation of empathy and distress on the EMP track.

1 Introduction

This paper focuses on Track 3: Empathy Prediction (EMP) of the WASSA 2024 Shared Task 1 at ACL 2024 (Giorgi et al., 2024). This task aims to predict empathy and distress based on essays. Previous NLP research studied how empathy and distress are expressed in the text (Sedoc et al., 2019) and tried to predict the level of empathy and distress with computational methods (Buechel et al., 2018; Barriere et al., 2023). Predicting empathy and distress is an important task that can be applied to diverse contexts, including discerning empathetic conversation (Omitaomu et al., 2022).

Recently, researchers have started to use LLMs to detect psychological status based on text data. For example, Xu et al. (2024) tested multiple LLMs with different methods for the prediction tasks for stress, depression, and other mental states. LLMs have also been used in emotion classification (Nedilko and Chu, 2023) and cognitive distortion classification task (Chen et al., 2023). Lastly, Hasan et al. (2024) used LLMs to convert numerical data into meaningful text and rephrase the text for predicting empathy.

Previous studies have shown that prompt engineering can achieve promising results in predicting mental health. For example, Qin et al. (2023) used the Chain-of-Thought technique and clinically established diagnostic criteria (DSM) in prompt to predict depression on social media texts, showing the best performance across various settings. Chen et al. (2023) proposed Diagnose of Prompt based on cognitive psychology and showed the best performance in classifying cognitive distortions. These findings show that constructing the prompt can be an important factor affecting the prediction performance.

However, the definitions of psychological constructs of LLMs could not be the same as the task defined. For example, LLMs could have different nuances for empathy and distress compared to the original research (Buechel et al., 2018). Figure 1 shows the definitions of empathy and distress of ChatGPT (gpt-3.5-turbo) and from the original paper (Buechel et al., 2018). While they shared the general meanings of the two constructs, ChatGPT's responses don't have the detailed nuances defined in the original paper. If the LLMs have different notions for the target variables, the prediction performance could be worse compared to having aligned definitions of psychological constructs.

In addition, empathy and distress could be ex-

Proceedings of the 14th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis, pages 380–384

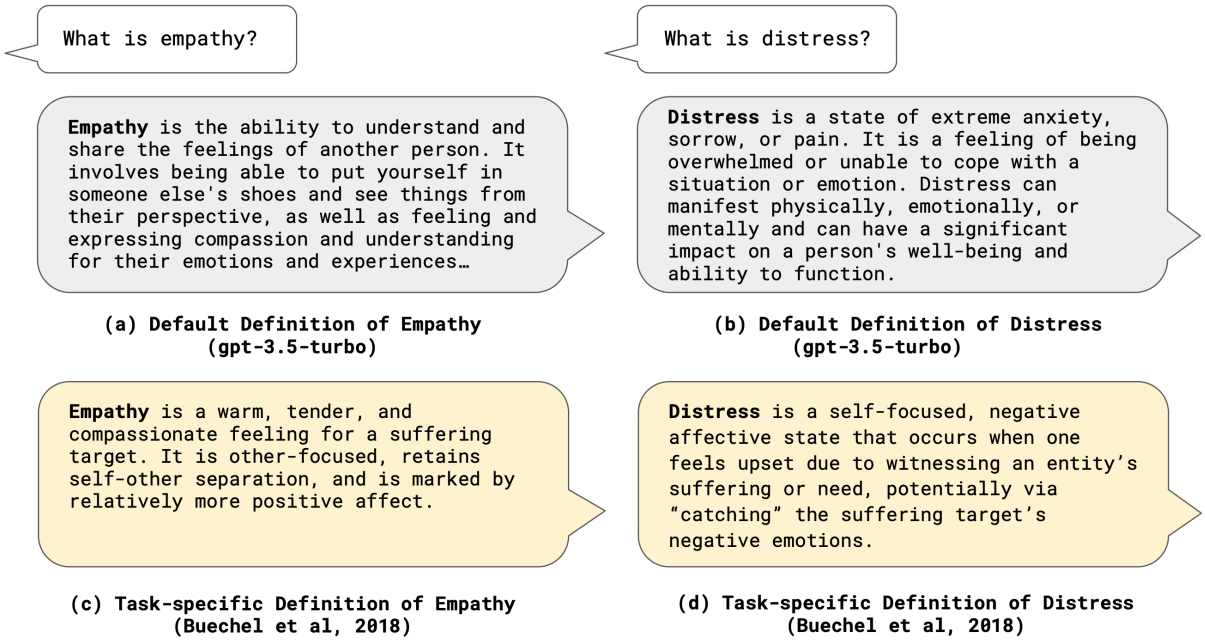

Figure 1: The Definitions of Empathy and Distress of ChatGPT and The Original Paper (Buechel et al., 2018)

pressed differently depending on the situation. This is due to the context-dependent nature of psychological constructs (Demszky et al., 2023). In other words, the way people express empathy could differ depending on who and how they communicate. Thus, it can be helpful to offer the detailed context of the dataset to LLMs.

To address the problems mentioned above, this study proposes a new prompt strategy, called Task-Aligned Prompt (TAP), for predicting empathy and distress. TAP consists of 1) definition alignment and 2) the context of the dataset. We tested the performance of our prompt compared to a plain prompt across zero-shot and few-shot settings with the models ChatGPT and GPT-4o. Our results show that the TAP-ChatGPT-zero-shot model achieved the best average Pearson correlation of empathy and distress on the development set. Our final submission ranked top 1 showing the best average Pearson correlation on the EMP track. This study shows the potential efficiency of LLMs in empathy and distress prediction and the strength of our approach.

2 Dataset

The dataset for Track 3 (EMP) consists of the level of empathy and distress of the writers, their essays, and the index of news articles (Buechel et al., 2018). The scores of both empathy and distress were measured with a 7-Likert scale using Batson's Empathic Concern – Personal Distress Scale (Batson et al., 1987). Thus, the level of empathy and distress range from 1 (not at all) to 7 (extremely).

Table 1 shows the number of instances for each dataset. The training set was only used in the few-shot prompting to give examples. The development dataset was used for both zero and few-shot promptings. The final evaluation was conducted on the test set.

Dataset	Instances
Train	1000
Dev	63
Test	83

Table 1: The Statistics of the Dataset

3 Methods

Our proposed Task-Aligned Prompt (TAP) aims to align LLMs for task-specific purposes. It mainly consists of 1) definition alignment and 2) dataset alignment. In the first stage, the prompts start with the definition of empathy and distress for each prediction task. The definitions of distress and empathy are retrieved from the previous paper that collected the dataset (Buechel et al., 2018) and the scales used to measure these states (Batson et al., 1987). In the second stage, the context of the text data was also retrieved from the original

Model	Average Pearson Correlation		Empathy Pearson Correlation		Distress Pearson Correlation	
	Task-Aligned	Plain	Task-Aligned	Plain	Task-Aligned	Plain
ChatGPT zero-shot	**0.511** ↑	0.494	$0.610_{0.010}$	$\mathbf{0.682}_{0.011}$	$0.413_{0.037}$	$0.306_{0.004}$
ChatGPT one-shot	0.464 ↓	0.493	$0.569_{0.015}$	$0.639_{0.033}$	$0.360_{0.085}$	$0.347_{0.109}$
ChatGPT three-shot	0.468 ↑	0.465	$0.571_{0.020}$	$0.593_{0.016}$	$0.365_{0.077}$	$0.337_{0.055}$
GPT-4o zero-shot	0.482 ↑	0.436	$0.520_{0.014}$	$0.439_{0.008}$	$0.445_{0.005}$	$0.433_{0.005}$
GPT-4o one-shot	0.492 ↑	0.468	$0.511_{0.016}$	$0.512_{0.059}$	$\mathbf{0.474}_{0.046}$	$0.424_{0.056}$
GPT-4o three-shot	0.484 ↑	0.477	$0.519_{0.051}$	$0.493_{0.030}$	$0.448_{0.032}$	$0.461_{0.044}$

Table 2: The Experiment Results for The Development Set

paper (Buechel et al., 2018). These prompts were included in the system prompts. Then, the prompts for task explanation, the constraint for output, and the target text are included. These three components are used as a plain prompt in this study. The details of our prompts are described in the Appendix A.

This study tested the TAP with zero-shot and few-shot prompting strategies. The prompts for zero-shot and few-shot prompting are the same except the few-shot prompting includes several examples (1 or 3), which are a pair of text and an answer (the level of empathy or distress). The examples used for few-shot prompting were chosen randomly in the training set.

For the experiments, we used the two models, ChatGPT (gpt-3.5-turbo) and the latest released GPT-4o model (gpt-4o) from OpenAI API [1]. We set the temperature to 0 and top_p to 1 which are the common practice for greedy decoding (Gupta et al., 2023). However, there are still variations in the responses across different runs. Thus, we ran each prompt three times and reported the average Pearson correlations of those three attempts for the development set.

For the final evaluation, the TAP-ChatGPT-zero-shot model was used since it performed the best on the development set. We tried to submit the results once for the final evaluation. Thus, we ran the test dataset 3 times, averaged the predicted values of the three results, and submitted those values. Our submission achieved the best score on the EMP track.

4 Results

Table 2 shows the results of our experiments for the development set. It shows the average, empathy, and distress Pearson correlations for the plain and our proposed prompts for each model and strategy. The average Pearson correlation means the average of empathy and distress Pearson correlation. For each empathy and distress Pearson correlation, we reported the average values and the standard deviation of performances across 3 runs for all cases. The best performances for each Pearson correlation were highlighted.

The TAP-ChatGPT-zero-shot model performed the best, showing the highest average Pearson correlation (r = .51). The Plain-ChatGPT-zero-shot model showed the best performance for the empathy Pearson correlation (r = .49) while the TAP-GPT4o-one-shot model performs the best for the distress prediction (r = 0.47). We marked the arrows next to TAP performances on the average Pearson correlation. The green arrow means the performance of our prompt is better than the plain prompt with the same models and strategies. Conversely, the red arrow means the plain prompt performed better than our prompt. We found that models with TAP outperformed every case except for one case.

Lastly, the TAP-ChatGPT-zero-shot model was used for the final evaluation. Table 3 shows the results of the top 3 teams in the EMP track. Our team, RU, ranked in the top 1 with an average Pearson coefficient of 0.453. Specifically, the Pearson coefficient for empathy and distress of our submission was 0.523 and 0.383 respectively.

Rank	Team Name	Score
1	**RU (Ours)**	**0.453**
2	Chinchunmei	0.393
3	FraunhoferSIT	0.385

Table 3: The Performances of The Top 3 Teams of EMP Track

[1] https://platform.openai.com/docs/models

5 Conclusion

This study showed the potential of LLMs in empathy and distress prediction tasks and our proposed prompt, Task-Aligned Prompt. Our experimental results showed that constructing prompts for LLMs to align the definitions of empathy and distress to the task and offering context of the dataset can benefit the prediction performance. Particularly, the TAP-ChatGPT-zero-shot model showed the best average Pearson correlation performance on the EMP track. These promising results strengthen the idea that the LLMs can be useful for predicting psychological states.

The limitation of our approach lies in its generalizability. Our approach may not be the most effective method for predicting empathy and distress across multiple datasets collected from diverse contexts and backgrounds. This is because our approach emphasizes aligning the detailed nuances and contexts to specific tasks.

Future research can continue to find an efficient prompting strategy for predicting empathy and distress. As mentioned above, researchers can study the one-size-fits-all prompts that can be applied to multiple datasets. In addition, improving the interpretability of LLMs prediction is also an important task in this field (Qin et al., 2023; Yang et al., 2024). While this study only focuses on predicting empathy and distress using LLMs, future studies can consider adding more layers to enhance the explainability and interpretability of LLMs.

References

Valentin Barriere, João Sedoc, Shabnam Tafreshi, and Salvatore Giorgi. 2023. Findings of wassa 2023 shared task on empathy, emotion and personality detection in conversation and reactions to news articles. In *Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, pages 511–525.

C Daniel Batson, Jim Fultz, and Patricia A Schoenrade. 1987. Distress and empathy: Two qualitatively distinct vicarious emotions with different motivational consequences. *Journal of personality*, 55(1):19–39.

Sven Buechel, Anneke Buffone, Barry Slaff, Lyle Ungar, and João Sedoc. 2018. Modeling empathy and distress in reaction to news stories. *arXiv preprint arXiv:1808.10399*.

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

Dorottya Demszky, David Yang, David S. Yeager, and et al. 2023. Using large language models in psychology. *Nature Reviews Psychology*, 2:688–701.

Salvatore Giorgi, João Sedoc, Valentin Barriere, and Shabnam Tafreshi. 2024. Findings of wassa 2024 shared task on empathy and personality detection in interactions. In *Proceedings of the 14th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*.

Shashank Gupta, Vaishnavi Shrivastava, Ameet Deshpande, Ashwin Kalyan, Peter Clark, Ashish Sabharwal, and Tushar Khot. 2023. Bias runs deep: Implicit reasoning biases in persona-assigned llms. *arXiv preprint arXiv:2311.04892*.

Md Rakibul Hasan, Md Zakir Hossain, Tom Gedeon, and Shafin Rahman. 2024. Llm-gem: Large language model-guided prediction of people's empathy levels towards newspaper article. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 2215–2231.

Andrew Nedilko and Yi Chu. 2023. Team bias busters at wassa 2023 empathy, emotion and personality shared task: Emotion detection with generative pretrained transformers. In *Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, pages 569–573.

Damilola Omitaomu, Shabnam Tafreshi, Tingting Liu, Sven Buechel, Chris Callison-Burch, Johannes Eichstaedt, Lyle Ungar, and João Sedoc. 2022. Empathic conversations: A multi-level dataset of contextualized conversations. *Preprint*, arXiv:2205.12698.

Wei Qin, Zetong Chen, Lei Wang, Yunshi Lan, Weijieying Ren, and Richang Hong. 2023. Read, diagnose and chat: Towards explainable and interactive llms-augmented depression detection in social media. *arXiv preprint arXiv:2305.05138*.

João Sedoc, Sven Buechel, Yehonathan Nachmany, Anneke Buffone, and Lyle Ungar. 2019. Learning word ratings for empathy and distress from document-level user responses. *arXiv preprint arXiv:1912.01079*.

Xuhai Xu, Bingsheng Yao, Yuanzhe Dong, Saadia Gabriel, Hong Yu, James Hendler, Marzyeh Ghassemi, Anind K Dey, and Dakuo Wang. 2024. Mental-llm: Leveraging large language models for mental health prediction via online text data. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 8(1):1–32.

Kailai Yang, Tianlin Zhang, Ziyan Kuang, Qianqian Xie, Jimin Huang, and Sophia Ananiadou. 2024. Mentallama: interpretable mental health analysis on social media with large language models. In *Proceedings of the ACM on Web Conference 2024*, pages 4489–4500.

A Prompts

The details of our prompts are described in Table 4. Specifically, the prompts were made as follows:

- Task-Aligned Prompt = Definition (Empathy or Distress) + Dataset Context + Task + Output Constraint + Text
- Plain Prompt = Task + Output Constraint + Text

The definition and the context of the dataset were written in the system prompt. The rest of the components were written in the user prompt. For few-shot prompting, we gave a pair of text and response sets in the form of Text: [text], and Response: [response] in the Text section of the prompt.

Name	Prompt
Definition (Empathy; system)	Empathy is a warm, tender, and compassionate feeling for a suffering target. It is other-focused, retains self-other separation, and is marked by relatively more positive affect. Empathy consists of warm, tender, sympathetic, softhearted, moved, and compassionate feelings.
Definition (Distress; system)	Distress is a self-focused, negative affective state that occurs when one feels upset due to witnessing an entity's suffering or need, potentially via "catching" the suffering target's negative emotions. Distress consists of worried, upset, troubled, perturbed, grieved, disturbed, alarmed, and distressed feelings.
Dataset context (system)	The following text is the reactions of people after reading news articles. They shared their feelings as they would with a friend in a private message or with a group of friends as a social media post.
Task	Evaluate the level of [empathy or distress] of the writer who wrote this text.
Output constraint	The answer should only contain a float value ranging from 1.0 (not at all) to 7.0 (extremely) using three decimal places.
Text	Text: [text]

Table 4: Prompt Design