Enhanced Detection of Conversational Mental Manipulation Through Advanced Prompting Techniques

Ivory Yang[1] **Xiaobo Guo**[2] **Sean Xie**[3] **Soroush Vosoughi**[4]
[1,2,3,4]Department of Computer Science, Dartmouth College
[1]Ivory.Yang.GR@dartmouth.edu
[4]Soroush.Vosoughi@dartmouth.edu

Abstract

This study presents a comprehensive, long-term project to explore the effectiveness of various prompting techniques in detecting dialogical mental manipulation. We implement Chain-of-Thought prompting with Zero-Shot and Few-Shot settings on a binary mental manipulation detection task, building upon existing work conducted with Zero-Shot and Few-Shot prompting. Our primary objective is to decipher why certain prompting techniques display superior performance so as to craft a novel framework tailored for the detection of mental manipulation. Preliminary findings suggest that advanced prompting techniques may not be suitable for more complex models if they are not trained through example-based learning.

1 Introduction

Mental manipulation is a subtle form of psychological influence on preferences and choice (Barnhill, 2014; Bublitz and Merkel, 2014), and its impact on society has been exacerbated by advancements in technology (Carroll et al., 2023). The detection of such manipulative language poses a significant hurdle within the field of Natural Language Processing (NLP) (Huffaker et al., 2020) due to its subtle, context-dependent, and inherently nuanced nature. To address these challenges, we investigate the effectiveness of advanced prompting techniques in detecting mental manipulation. The results of our experiments show that while Chain-of-Thought (CoT) prompting achieves exceptional accuracy across scenarios, special attention must be given to the learning configuration in order to ensure its high performance [1].

2 Related Work

Wang et al. (2024) proposed MentalManip, a dataset comprising of over 4000 human-annotated dialogues focused on mental manipulation. By leveraging Zero-Shot and Few-Shot prompting, the authors set a foundational understanding of model limitations in scenarios devoid of explicit toxic indicators. In recent years, there has been an advent in sophisticated prompting techniques to improve LLM performance on complex NLP tasks. Chain-of-Thought prompting (Wei et al., 2022) uses example-based learning for models to generate intermediate reasoning steps before arriving at a conclusion. As an extension, Zero-Shot Chain-of-Thought (Kojima et al., 2022) relies solely on crafted prompts that guide the model to articulate its gradiational reasoning process. This paper aims to apply these advanced prompting techniques to the problem of mental manipulation detection.

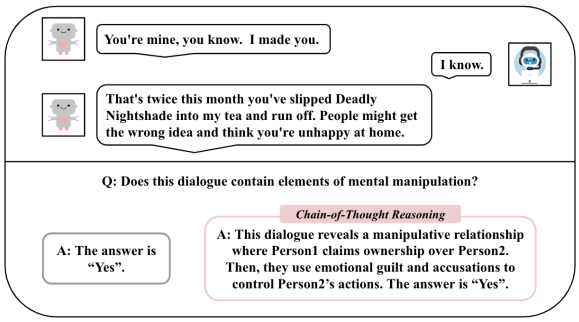

Figure 1: An example dialogue incorporating Chain-of-Thought reasoning in mental manipulation detection

3 Experiment

3.1 Setup

We conducted our experiments using the **MentalManipCon** dataset (Wang et al., 2024), a rigorously selected corpus of dialogues on mental manipulation, with consensus agreement amongst all three reviewers. We then conducted our experiments across two LLMs, GPT-3.5 and GPT-4o, to assess their performance in detecting mental manipulation. Both models were evaluated using the

[1]The code and data for the experiments presented in this paper can be found here.

Experiment Setting	GPT-3.5					GPT-4o				
	P	R	Acc	F_1^{mi}	F_1^{ma}	P	R	Acc	F_1^{mi}	F_1^{ma}
Zero-Shot	.750	.827	.693	.693	.620	.741	.952	.739	.739	.617
Few-Shot	.789	.769	.702	.702	.659	.749	.980	.762	.762	.641
CoT (Zero)	.725	.951	.721	.721	.579	.729	.950	.724	.724	.586
CoT (Few)	.714	.909	**.722**	.722	.673	.769	.909	**.778**	.778	.750

Table 1: Results of manipulation detection task on MentalManipCon dataset. P, R, Acc, F_1^{mi}, and F_1^{ma} stands for binary precision, binary recall, accuracy, micro F_1, and macro F_1 respectively. "CoT (Zero)": "Chain-of-Thought with Zero-Shot learning settings", "CoT (Few)": "Chain-of-Thought with Few-Shot learning settings".

same set of tasks, with effectiveness measured by Precision, Recall, Accuracy, and F1-Score across prompting strategies. Due to binary classification, Accuracy and Micro F1 reflect the same scores.

3.2 Prompting Techniques

Our study assesses the effectiveness of four different prompting strategies in identifying mental manipulation: Zero-Shot, Few-Shot, and CoT with Zero-Shot and Few-Shot settings. The Zero-Shot approach was implemented without providing any prior examples of the model. In contrast, the Few-Shot approach utilized a set of three examples (two manipulatives, one non-manipulative) randomly chosen to guide the model. The CoT strategy with Zero-Shot settings involved a modified prompt that encouraged the model to process its thoughts step-by-step, integrating a reasoning component into its responses. For CoT with Few-Shot settings, we enlisted two college students, both native English speakers, to manually annotate detailed, step-by-step reasoning for a randomly selected set of 42 examples. This annotated dataset was then employed to train our model for CoT prompting.

4 Results

From the results in Table 1, it is evident that Few-Shot CoT produces the best performance in terms of Accuracy, at 0.722 for GPT-3.5 and 0.778 for GPT-4o, which aligns with our expectations. However, it is notable that when upgrading from GPT-3.5 to GPT-4o, Zero-Shot CoT drops from second best performing to the worst performing technique, even worse than regular Zero-Shot. It also produces a significant number of false positives, as reflected by its Macro F1 scores, which are the lowest across all techniques for both GPT-3.5 and GPT-4o.

The exceptional performance of Few-Shot CoT can be attributed to its structured reasoning process, which closely mirrors human cognitive strategies. However, by comparing the performance of Zero-Shot settings, we observe that although CoT generally performs better on the task of mental manipulation, it should be combined with samples for learning. Considering the reduced Precision on GPT-3.5 and GPT-4o with Zero-Shot CoT, we assume that both models may wrongly understand the definition of mental manipulation, which is further enhanced during CoT. Therefore, we conduct a pilot check on the generated reasons.

After manually checking the results, GPT-4o places an overemphasis on verbal cues and misinterprets fragmented or informal speech. The model attributes manipulation to communication style rather than actual manipulative intent. GPT-4o also appears to be biased towards conflict, detecting manipulation even in benign situations and interpreting neutral or vague responses as signs of manipulation. For mental manipulation detection, CoT **without example-based learning** may **perform worse** in relation to simpler techniques as **model complexity increases**.

5 Future Work

We aim to extend our analysis of CoT performance concerning model complexity and the definition of mental manipulation provided to the model. Moreover, we will explore other prompting techniques such as Iterative prompting (Wang et al., 2022a), Self-Consistency (Wang et al., 2022b) and Tree-of-Thoughts prompting (Long, 2023; Yao et al., 2024). We also seek to take into account serial position effects (Guo and Vosoughi, 2024) to analyze how the placement of information within prompts affects detection accuracy. Given that this is a long-term project on mental manipulation detection, we may potentially expand our scope to consider the impact of gender (Grieve et al., 2019), stereotypes (Ma et al., 2023), and biases (Xie et al., 2024) in our results.

References

Anne Barnhill. 2014. What is manipulation? In *Manipulation: Theory and Practice*. Oxford University Press.

Jan Christoph Bublitz and Reinhard Merkel. 2014. Crimes against minds: On mental manipulations, harms and a human right to mental self-determination. *Criminal Law and Philosophy*, 8:51–77.

Micah Carroll, Alan Chan, Henry Ashton, and David Krueger. 2023. Characterizing manipulation from ai systems. In *Proceedings of the 3rd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*, pages 1–13.

Rachel Grieve, Evita March, and George Van Doorn. 2019. Masculinity might be more toxic than we think: The influence of gender roles on trait emotional manipulation. *Personality and Individual Differences*, 138:157–162.

Xiaobo Guo and Soroush Vosoughi. 2024. Serial position effects of large language models. *arXiv preprint arXiv:2406.15981*.

Jordan S Huffaker, Jonathan K Kummerfeld, Walter S Lasecki, and Mark S Ackerman. 2020. Crowdsourced detection of emotionally manipulative language. In *Proceedings of the 2020 CHI conference on human factors in computing systems*, pages 1–14.

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.

Jieyi Long. 2023. Large language model guided tree-of-thought. *arXiv preprint arXiv:2305.08291*.

Weicheng Ma, Henry Scheible, Brian Wang, Goutham Veeramachaneni, Pratim Chowdhary, Alan Sun, Andrew Koulogeorge, Lili Wang, Diyi Yang, and Soroush Vosoughi. 2023. Deciphering stereotypes in pre-trained language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 11328–11345.

Boshi Wang, Xiang Deng, and Huan Sun. 2022a. Iteratively prompt pre-trained language models for chain of thought. *arXiv preprint arXiv:2203.08383*.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2022b. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*.

Yuxin Wang, Ivory Yang, Saeed Hassanpour, and Soroush Vosoughi. 2024. MentalManip: A dataset for fine-grained analysis of mental manipulation in conversations. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3747–3764.

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.

Sean Xie, Saeed Hassanpour, and Soroush Vosoughi. 2024. Addressing healthcare-related racial and lgbtq+ biases in pretrained language models. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 4451–4464.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. 2024. Tree of thoughts: Deliberate problem solving with large language models. *Advances in Neural Information Processing Systems*, 36.