Detecting Conversational Mental Manipulation with Intent-Aware Prompting

Jiayuan Ma[♠*], **Hongbin Na**[♡*], **Zimu Wang**[♣], **Yining Hua**[△], **Yue Liu**[◇], **Wei Wang**[♣], **Ling Chen**[♡]

[♠]The University of Sydney [♡]University of Technology Sydney
[♣]Xi'an Jiaotong-Liverpool University [△]Harvard University
[◇]University of New South Wales

jima3429@uni.sydney.edu.au, hongbin.na@student.uts.edu.au
zimu.wang19@student.xjtlu.edu.cn, yininghua@g.harvard.edu
z5472597@ad.unsw.edu.au, wei.wang03@xjtlu.edu.cn, ling.chen@uts.edu.au

Abstract

Mental manipulation severely undermines mental wellness by covertly and negatively distorting decision-making. While there is an increasing interest in mental health care within the natural language processing community, progress in tackling manipulation remains limited due to the complexity of detecting subtle, covert tactics in conversations. In this paper, we propose Intent-Aware Prompting (IAP), a novel approach for detecting mental manipulations using large language models (LLMs), providing a deeper understanding of manipulative tactics by capturing the underlying intents of participants. Experimental results on the MentalManip dataset demonstrate superior effectiveness of IAP against other advanced prompting strategies. Notably, our approach substantially reduces false negatives, helping detect more instances of mental manipulation with minimal misjudgment of positive cases. The code of this paper is available at https://github.com/Anton-Jiayuan-MA/Manip-IAP.

1 Introduction

Human interactions inevitably involve varying degrees of mutual influence, from ethical persuasion based on facts to more harmful tactics like coercion and manipulation (Fischer, 2022). Manipulation represents a serious concern, as it involves the deliberate control or distortion of an individual's thoughts and emotions for personal gain (Barnhill, 2014). Such manipulation can lead to detrimental mental health issues if left unchecked. The ability to detect and address these behaviors swiftly and accurately is critical for protecting individuals from potential mental health deterioration and ensuring their well-being.

Large language models (LLMs), known for their exceptional capability to process and reason over lengthy contexts (Peng et al., 2023), are ideally suited for detecting mental manipulations. Recent studies have shown the reliability of LLMs in addressing mental health issues (Hua et al., 2024; Na, 2024; Na et al., 2024). One prominent research direction involves leveraging prompt engineering techniques, such as zero-shot, few-shot, chain-of-thought (CoT), and diagnosis-of-thought (DoT) prompting (Chen et al., 2023; Schulhoff et al., 2024), or fine-tuning the models on curated, annotated datasets sourced from social media platforms like Reddit and Twitter (Wang et al., 2024a; Yang et al., 2024b; Qian et al., 2024).

To advance the analysis of manipulative dialogues, Wang et al. (2024b) introduces the first dataset, MentalManip, specialized for mental manipulation detection and classification. Despite their strengths, LLMs exhibit notable difficulties in identifying manipulative dialogues; in particular, the false negative rate is almost twice the false positive rate, as evidenced by our pilot study (see Section 2). This limitation poses a substantial problem for real-world applications, where early detection of mental manipulation is critical. Building on this, Yang et al. (2024a) concludes that a combination of few-shot and CoT prompting significantly enhances performance, highlighting the necessity for more advanced prompting techniques to improve LLM performance in this challenging task.

In response, we propose **Intent-Aware Prompting (IAP)**, a novel approach to enhance LLM's Theory of Mind (ToM) and its ability in detecting mental manipulations from dialogues. As shown in Figure 2, IAP leverages a distinct analysis of the underlying intents of both participants in the conversation, providing a deeper understanding of manipulative tactics. We perform extensive experiments on the MentalManip dataset, showcasing the superior effectiveness of IAP against other advanced prompting techniques, such as few-shot and

[*]Equal contribution.

Proceedings of the 31st International Conference on Computational Linguistics, pages 9176–9183
January 19–24, 2025.

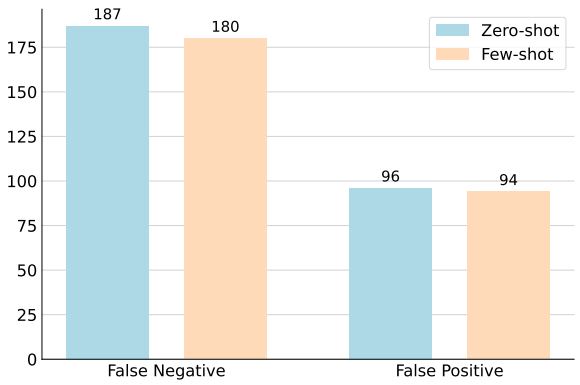

Figure 1: Comparison of false negatives and false positives in mental manipulation detection using zero-shot and few-shot prompting.

CoT prompting. Notably, IAP significantly reduces the false negative rate, highlighting its practical relevance for early detection of mental manipulation in real-world applications.

Our key contributions are as follows: (1) the introduction of IAP for detecting mental manipulation in dialogues. It improves the ToM of LLMs via intent summarization, thus improving model performance on the task; (2) extensive experiments on the MentalManip dataset, which demonstrates that IAP outperforms baseline methods and substantially reduces false negatives; (3) human evaluation of the intent summarization process, confirming the high quality of the generated intents.

2 Observation

In pilot experiments using the zero-shot approach to detect mental manipulation (Wang et al., 2024b), we observe that the false negative (FN) rate is approximately **double** that of the false positive (FP) rate. This indicates challenges with the model's ability to recognize manipulation patterns or insufficient feature representation in the input data. This observation aligns with the reality that mental manipulation is inherently difficult to detect, even for humans, due to its subtle and covert nature (Barnhill, 2014). While Wang et al. (2024b) have also attempted to improve mental manipulation detection using the few-shot approach, the challenge of performance imbalance remains unresolved. The changes in FNs and FPs between the zero-shot and few-shot methods are illustrated in Figure 1.

3 Methodology

Psychological research suggests that individuals with strong Theory of Mind (ToM) are more adept

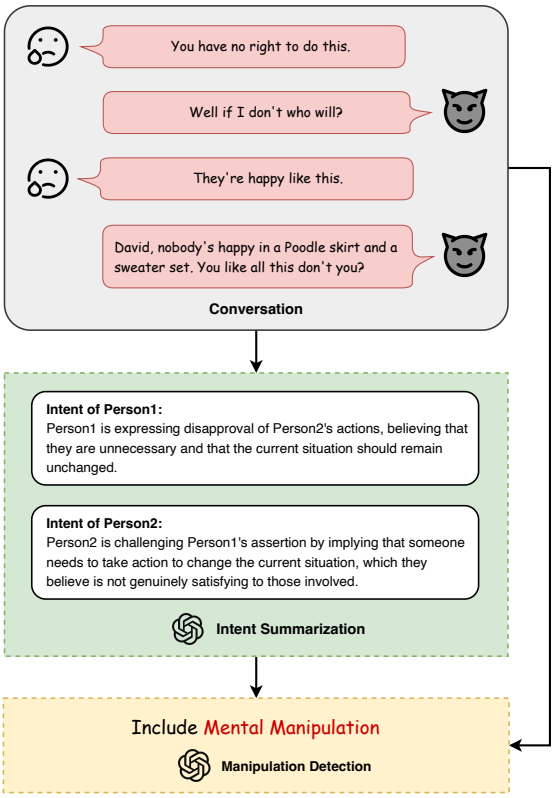

Figure 2: Overall framework of Intent-Aware Prompting (IAP) on mental manipulation detection.

at discerning subtle differences in others' intentions (Byom and Mutlu, 2013). Conversely, those with ToM deficits are more susceptible to manipulations (Kern et al., 2009; Lampron et al., 2024). LLMs have been proven to improve ToM task performance with CoT reasoning, enhancing their ability to infer complex social cues and mental states (Moghaddam and Honey, 2023; Chen et al., 2024).

Building on these findings, we propose IAP for improving the ToM of LLMs and their capability in detecting mental manipulations. By incorporating intent-based reasoning, our goal is to tackle the high FN rate observed in our pilot experiments (§2), by improving model ability to detect subtle manipulative behavior that might be neglected. In this section, we present two key components for implementing IAP – Intent Summarization (§3.1) and Manipulation Detection (§3.2). Figure 2 shows an overview of Intent-Aware Prompting.

3.1 Intent Summarization

Consider we have a conversation $\mathbb{D}$, structured as:

$$\mathbb{D} = \{\boldsymbol{u}_{A1}, \boldsymbol{u}_{B1}, \ldots, \boldsymbol{u}_{An}, \boldsymbol{u}_{Bn}\},$$

where u_A represents utterances by Person A and u_B represents utterances by Person B. We design

Method	FN↓		FP↓		Accuracy↑		Precision↑		Recall↑		$F1_{Weighted}$↑		$F1_{Macro}$ ↑	
Zero-Shot	187	-	96	-	0.677	-	0.813	-	0.691	-	0.687	-	0.649	-
Few-Shot	180	-3.7%	**94**	-2.1%	0.687	1.5%	**0.819**	0.7%	0.702	1.6%	0.696	1.3%	0.659	1.5%
Zero-Shot CoT	159	-15.0%	101	5.2%	0.703	3.8%	0.815	0.2%	0.737	6.7%	0.710	3.3%	0.670	3.2%
Intent-Aware	**130**	-30.5%	110	14.6%	**0.726**	7.2%	0.812	-0.1%	**0.785**	13.6%	**0.728**	6.0%	**0.685**	5.5%

Table 1: **Result of detecting mental manipulation using GPT-4.** Metrics with an upward arrow ↑ indicate higher values are better, while metrics with a downward arrow ↓ indicate lower values are better. Using zero-shot as comparison, darker green means better performance, and darker red means worse performance of the model.

an intent summarization prompt $P_{\text{IS}}(\cdot)$, which consists of an intent summarization instruction for the two people $\boldsymbol{p}_A$ and $\boldsymbol{p}_B$. The intent summaries $\boldsymbol{i}_A$ and $\boldsymbol{i}_B$ can be defined as:

$$\boldsymbol{i}_A = \text{LLM}(\mathbb{D}, P_{\text{IS}}(\boldsymbol{p}_A)), \quad (1)$$

$$\boldsymbol{i}_B = \text{LLM}(\mathbb{D}, P_{\text{IS}}(\boldsymbol{p}_B)), \quad (2)$$

where $\text{LLM}(\cdot)$ represents an LLM used to generate the intent summary. The detailed prompt is provided in Appendix A.

Remark. The entire conversation $\mathbb{D}$ is used instead of just the utterances from one individual because understanding each person's intent relies on a holistic understanding of the contexts.

3.2 Manipulation Detection

Given the conversation $\mathbb{D}$, and the intent summaries $\boldsymbol{i}_A$ and $\boldsymbol{i}_B$ calculated from $\mathbb{D}$ using Equations 1 and 2, the mental manipulation detection process can be defined as:

$$r = \text{LLM}(\mathbb{D}, \boldsymbol{i}_A, \boldsymbol{i}_B, P_{\text{MD}}), \quad (3)$$

where r denotes the detection result, with $r \in \{0, 1\}$. Specifically, $r = 0$ means that no mental manipulation has been detected, while $r = 1$ means that mental manipulation is present. The function $\text{LLM}(\cdot)$ refers to a LLM used to process the full conversation $\mathbb{D}$ along with the intent summaries $\boldsymbol{i}_A$ and $\boldsymbol{i}_B$ and the manipulation detection prompt P_{MD}. The prompt P_{MD} is specifically designed to evaluate the interaction between the two intent summaries and detect potential manipulation behaviors in the conversation. The detailed prompt is provided in Appendix A.

4 Experiments

4.1 Experimental Settings

Dataset. We conducted experiments using the MentalManip dataset (Wang et al., 2024b), which provides multi-level annotations aimed at detecting and classifying mental manipulations. It consists of 4,000 multi-turn fictional dialogues between two characters derived from online movie scripts and includes annotation across three dimensions: the presence of manipulation, manipulation technique, and targeted vulnerability. For our experiments, we sampled a subset with 30% instances (1.5 times of the original test set) in $\text{MentalManip}_{\text{con}}$, a subset of MentalManip with full annotator consensus, ensuring high quality and consistency in the experimented data.

Evaluation Metrics. Following Wang et al. (2024b), we evaluated performance using accuracy, precision, recall, and F1-score (weighted and macro). Furthermore, the analysis of the false predictions, including false negatives (FN) and false positives (FP), is also crucial, as it provides insights on how well the method works to identify psychologically manipulated dialogues without excessively exaggerating the presence of the positive class.

Baselines. In accordance with the previous work, we compared the performance of IAP against the following baselines: (1) **Zero-shot prompting**, which enables LLMs to perform the task based solely on the given input. (2) **Few-shot prompting** (Brown et al., 2020), which generalizes LLMs to the task by providing a few examples within the input prompt. We randomly selected three examples from the data subset not used for testing, with a proportion of 1:2 of manipulative to non-manipulative. (3) **Chain-of-Thought (CoT) prompting** (Kojima et al., 2022), which enhances LLM's reasoning capabilities by generating intermediate reasoning steps within its output, enabling more complicated problem-solving and decision-making processes.

4.2 Experimental Results

Main Results. Table 1 illustrates the experimental results of IAP against baselines, in which GPT-

Rating Category	Percentage
Accurate	82%
Inaccurate	18%

Table 2: Percentage of intents rated as accurate and inaccurate based on human evaluation.

4[1] (OpenAI, 2024) was utilized in all experiments. From the table, we observed the effectiveness of IAP on mental manipulation detection by achieving the best performance on nearly all evaluation metrics, demonstrating substantial improvements in accuracy (+7.2%), recall (+13.6%), weighted F1-score (+6.0%), and macro F1-score (+5.5%) compared with zero-shot prompting. This underscores the potent efficacy of analyzing speakers' intentions in identifying the existence of mental manipulation within dialogues and the improvement of the ToM of the model. Besides, IAP achieved the lowest number of false negatives (130), representing a 30.5% reduction compared to zero-shot prompting and substantially outperforming other baseline methods, reinforcing its ability to detect a higher number of mental manipulative dialogues. While there is a trade-off with an increase in false positives (+14.6%), the substantial reduction in false negatives is far more critical, where early detection and intervention for potential mental health concerns are much more paramount.

Human Evaluation of Generated Intents. To assess the quality of the generated intents in the absence of references, we conducted a human evaluation to verify whether they correctly identified the manipulators. We selected 50 dialogues from the dataset that exhibited mental manipulations, and two annotators independently assessed each dialogue, labeling Person A, Person B, or both as the manipulator(s). The inter-annotator agreement reached 74%, and the discrepancies were resolved through discussion to reach a consensus. During evaluation, we verified if the generated intents accurately pointed to the labeled manipulator(s). As shown in Table 2, 82% of the intents correctly identified the manipulator(s), demonstrating the capability of IAP in producing high-quality intents that significantly aid in the detection of manipulations. Some examples are in Appendix B.

[1] `gpt-4-1106-preview`

5 Related Work

Mental Manipulation Detection. Dialogue-based classification poses unique challenges due to the dynamic, multi-turn nature of conversations. These challenges include handling long text sequences, managing context shifts, capturing speaker roles and intents, and modeling nuanced interactions across multiple turns. In mental healthcare, dialogue-based classification has been primarily used for identifying mental health conditions (Hua et al., 2024) and detecting toxic behaviors (Ozoh et al., 2019), including threats, obscenity, insults, identity-based hate, harassment, and socially disruptive persuasion (Sheth et al., 2022).

However, the research on mental manipulation remains underexplored. The only existing work (Yang et al., 2024a) investigates prompting techniques for mental manipulation detection with limited standard prompting techniques. Different from the previous work, we introduce a novel approach for by analyzing underlying intents of both participants in the conversation, offering a deeper understanding of manipulative tactics.

LLMs and Theory of Mind. LLMs with exceptional capability to process and reason over lengthy contexts have become the cornerstone of numerous NLP tasks (Peng et al., 2023; Wang et al., 2024c), making them particularly well-suited for dialogue-based applications (Na et al., 2024; Lee et al., 2024; Zheng et al., 2024a,b; Iftikhar et al., 2024), which require comprehending not only individual turns but also the evolution of context, tone, and intent throughout the conversations.

"Theory of Mind" (ToM) refers to the ability to infer and understand the mental states, intentions, beliefs, and emotions of others. While traditionally regarded a human cognitive trait, recent research suggests that LLMs can simulate aspects of this ability, even surpassing humans in tasks like recognizing irony and false beliefs (Strachan et al., 2024). This capability is particularly valuable for mental manipulation detection, where accurately interpreting and predicting speakers' intentions and emotional states is essential for uncovering manipulative strategies in dialogues (Kern et al., 2009; Lampron et al., 2024).

6 Conclusion

We introduced Intent-Aware Prompting (IAP), a novel approach to enhance LLM's ability in de-

tecting mental manipulations from dialogues. It enhanced LLM's ToM and its manipulation detection capability by distinctly analyzing the underlying intents of both participants, offering a more nuanced understanding of manipulative strategies. Through comprehensive experiments on the MentalManip dataset, IAP consistently outperformed other advanced prompting techniques, such as few-shot and CoT prompting, across multiple metrics. Notably, it achieved a substantial reduction in false negatives, a crucial improvement in the context of mental health support systems where early detection of psychological manipulation is key to timely interventions. In the future, we will expand IAP to broader mental health applications to more real-world scenarios.

Limitations

The limitations of this paper are as follows: (1) Although performance increased, the reduction in false negatives led to a slight increase in false positives. While its real-world impact is minimal, it might introduce therapeutic costs. Future research can focus on optimizing the trade-off between false negatives and false positives. (2) Due to only one dataset available, we only tested the performance of IAP on the MentalManip dataset. Future research can develop more diverse mental manipulation datasets encompassing both high-resource and low-resource languages and validate the generalizability of IAP across different linguistic and contextual settings.

References

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

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Lindsey J. Byom and Bilge Mutlu. 2013. Theory of mind: mechanisms, methods, and new directions. *Frontiers in Human Neuroscience*, 7.

Zhiyu Chen, Yujie Lu, and William Wang. 2023. Empowering psychotherapy with large language models: Cognitive distortion detection through diagnosis of thought prompting. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 4295–4304, Singapore. Association for Computational Linguistics.

Zhuang Chen, Jincenzi Wu, Jinfeng Zhou, Bosi Wen, Guanqun Bi, Gongyao Jiang, Yaru Cao, Mengting Hu, Yunghwei Lai, Zexuan Xiong, and Minlie Huang. 2024. ToMBench: Benchmarking theory of mind in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15959–15983, Bangkok, Thailand. Association for Computational Linguistics.

Alexander Fischer. 2022. Then again, what is manipulation? a broader view of a much-maligned concept. *Philosophical Explorations*, 25(2):170–188.

Yining Hua, Fenglin Liu, Kailai Yang, Zehan Li, Hongbin Na, Yi han Sheu, Peilin Zhou, Lauren V. Moran, Sophia Ananiadou, Andrew Beam, and John Torous. 2024. Large language models in mental health care: a scoping review. *Preprint*, arXiv:2401.02984.

Zainab Iftikhar, Sean Ransom, Amy Xiao, and Jeff Huang. 2024. Therapy as an nlp task: Psychologists' comparison of llms and human peers in cbt. *Preprint*, arXiv:2409.02244.

R. S. Kern, M. F. Green, A. P. Fiske, K. S. Kee, J. Lee, M. J. Sergi, W. P. Horan, K. L. Subotnik, C. A. Sugar, and K. H. Nuechterlein. 2009. Theory of mind deficits for processing counterfactual information in persons with chronic schizophrenia. *Psychological Medicine*, 39(4):645–654.

Takeshi Kojima, Shixiang (Shane) Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. In *Advances in Neural Information Processing Systems*, volume 35, pages 22199–22213. Curran Associates, Inc.

Mireille Lampron, Amélie M. Achim, Dominick Gamache, Allyson Bernier, Stéphane Sabourin, and Claudia Savard. 2024. Profiles of theory of mind impairments and personality in clinical and community samples: integrating the alternative dsm-5 model for personality disorders. *Frontiers in Psychiatry*, 14.

Suyeon Lee, Sunghwan Kim, Minju Kim, Dongjin Kang, Dongil Yang, Harim Kim, Minseok Kang, Dayi Jung, Min Hee Kim, Seungbeen Lee, Kyong-Mee Chung, Youngjae Yu, Dongha Lee, and Jinyoung Yeo. 2024. Cactus: Towards psychological counseling conversations using cognitive behavioral theory. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 14245–14274, Miami, Florida, USA. Association for Computational Linguistics.

Shima Rahimi Moghaddam and Christopher J. Honey. 2023. Boosting theory-of-mind performance in large language models via prompting. *Preprint*, arXiv:2304.11490.

Hongbin Na. 2024. CBT-LLM: A Chinese large language model for cognitive behavioral therapy-based mental health question answering. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 2930–2940, Torino, Italia. ELRA and ICCL.

Hongbin Na, Tao Shen, Shumao Yu, and Ling Chen. 2024. Multi-session client-centered treatment outcome evaluation in psychotherapy. *Preprint*, arXiv:2410.05824.

OpenAI. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

PA Ozoh, Adepeju Abeke Adigun, and MO Olayiwola. 2019. Identification and classification of toxic comments on social media using machine learning techniques. *International Journal of Research and Innovation in Applied Science (IJRIAS)*, 4(11):142–147.

Hao Peng, Xiaozhi Wang, Jianhui Chen, Weikai Li, Yunjia Qi, Zimu Wang, Zhili Wu, Kaisheng Zeng, Bin Xu, Lei Hou, and Juanzi Li. 2023. When does in-context learning fall short and why? a study on specification-heavy tasks. *Preprint*, arXiv:2311.08993.

Lu Qian, Yuqi Wang, Zimu Wang, Haiyang Zhang, Wei Wang, Ting Yu, and Anh Nguyen. 2024. Domain-specific guided summarization for mental health posts. *Preprint*, arXiv:2411.01485.

Sander Schulhoff, Michael Ilie, Nishant Balepur, Konstantine Kahadze, Amanda Liu, Chenglei Si, Yinheng Li, Aayush Gupta, HyoJung Han, Sevien Schulhoff, Pranav Sandeep Dulepet, Saurav Vidyadhara, Dayeon Ki, Sweta Agrawal, Chau Pham, Gerson Kroiz, Feileen Li, Hudson Tao, Ashay Srivastava, Hevander Da Costa, Saloni Gupta, Megan L. Rogers, Inna Goncearenco, Giuseppe Sarli, Igor Galynker, Denis Peskoff, Marine Carpuat, Jules White, Shyamal Anadkat, Alexander Hoyle, and Philip Resnik. 2024. The prompt report: A systematic survey of prompting techniques. *Preprint*, arXiv:2406.06608.

Amit Sheth, Valerie L. Shalin, and Ugur Kursuncu. 2022. Defining and detecting toxicity on social media: context and knowledge are key. *Neurocomputing*, 490:312–318.

James W. A. Strachan, Dalila Albergo, Giulia Borghini, Oriana Pansardi, Eugenio Scaliti, Saurabh Gupta, Krati Saxena, Alessandro Rufo, Stefano Panzeri, Guido Manzi, Michael S. A. Graziano, and Cristina Becchio. 2024. Testing theory of mind in large language models and humans. *Nature Human Behaviour*, 8(7):1285–1295.

Yuqi Wang, Zimu Wang, Nijia Han, Wei Wang, Qi Chen, Haiyang Zhang, Yushan Pan, and Anh Nguyen. 2024a. Knowledge distillation from monolingual to multilingual models for intelligent and interpretable multilingual emotion detection. In *Proceedings of the 14th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, pages 470–475, Bangkok, Thailand. Association for Computational Linguistics.

Yuxin Wang, Ivory Yang, Saeed Hassanpour, and Soroush Vosoughi. 2024b. 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, Bangkok, Thailand. Association for Computational Linguistics.

Zimu Wang, Lei Xia, Wei Wang Xjtlu, and Xinya Du. 2024c. Document-level causal relation extraction with knowledge-guided binary question answering. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 16944–16955, Miami, Florida, USA. Association for Computational Linguistics.

Ivory Yang, Xiaobo Guo, Sean Xie, and Soroush Vosoughi. 2024a. Enhanced detection of conversational mental manipulation through advanced prompting techniques. *Preprint*, arXiv:2408.07676.

Kailai Yang, Tianlin Zhang, Ziyan Kuang, Qianqian Xie, Jimin Huang, and Sophia Ananiadou. 2024b. Mentallama: Interpretable mental health analysis on social media with large language models. In *Proceedings of the ACM Web Conference 2024*, WWW '24, page 4489–4500, New York, NY, USA. Association for Computing Machinery.

Zhonghua Zheng, Lizi Liao, Yang Deng, Ee-Peng Lim, Minlie Huang, and Liqiang Nie. 2024a. Thoughts to target: Enhance planning for target-driven conversation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 21108–21124, Miami, Florida, USA. Association for Computational Linguistics.

Zhonghua Zheng, Lizi Liao, Yang Deng, Libo Qin, and Liqiang Nie. 2024b. Self-chats from large language models make small emotional support chatbot better. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11325–11345, Bangkok, Thailand. Association for Computational Linguistics.

A Prompts in Experiments

Zero-shot Prompting

```
I will provide you with a dialogue. Please determine if it contains elements of mental manipulation.
Just answer with 'Yes' or 'No', and don't add anything else.

<insert dialogue>
```

Few-shot Prompting

```
I will provide you with a dialogue. Please determine if it contains elements of mental manipulation.
Just answer with 'Yes' or 'No', and don't add anything else. Here are 3 examples:

Example 1:
<insert manipulative_dialogue1>
<insert manipulative_answer1>

Example 2:
<insert nonmanipulative_dialogue1>
<insert nonmanipulative_answer1>

Example 3:
<insert nonmanipulative_dialogue2>
<insert nonmanipulative_answer2>

<insert dialogue>
```

Zero-shot CoT Prompting

```
I will provide you with a dialogue. Please determine if it contains elements of mental manipulation.
Just answer with 'Yes' or 'No', and don't add anything else. Let's think step by step.

<insert dialogue>
```

Intent-Aware Prompting

```
I will provide you with a dialogue and intent of person1, and intent of person2. Please carefully
analyze the dialogue and intents, and determine if it contains elements of mental manipulation. Just
answer with 'Yes' or 'No', and don't add anything else.

<insert dialogue>
<insert person1's intent>
<insert person2's intent>
```

Figure 3: Examples of various prompting strategies used in the experiment: Zero-shot Prompting, Few-shot Prompting, Zero-shot CoT Prompting, and Intent-Aware Prompting. Each prompt is designed to assess the presence of mental manipulation in dialogues.

B Intent Summarization Examples

Intent Summarization Prompting

```
I will provide you with a dialogue. Please summarize the intent of the statement made by <insert person>
 in one sentence.

<insert dialogue>
```

Intent Summarization Example 1

```
Dialogue:
Person1: You have no right to do this.
Person2: Well if I don't who will?
Person1: They're happy like this.
Person2: David, nobody's happy in a Poodle skirt and a sweater set. You like all this don't you?
```
```
Person1's intent:
Person1 is expressing disapproval of Person2's actions, believing that they are unnecessary and that the
 current situation should remain unchanged.
```
```
Person2's intent:
Person2 is challenging Person1's assertion by implying that someone needs to take action to change the
current situation, which they believe is not genuinely satisfying to those involved.
```

Intent Summarization Example 2

```
Dialogue:
Person1: What are you doing here?
Person2: Nothing.  I just wanted you to know I was out.  I just wanted to see you.
Person1: Well, here I am.  See?
Person2: How are you doing?
Person1: George, you just can't show up, tell me you love me, and have everything be okay.
Person2: Dad.
Person1: What?
Person2: You can call me Dad if you want.
Person1: I don't want, alright?  It's not funny. I'm really pissed off, George.  You blew it, now leave
me alone.
Person2: Kristina, c'mon, I'm sorry.  I'm going to make this right.  I've got a few things going on...
Person1: What do you want from me?
Person2: Just to walk with you.  I want to be your dad again.
Person1: Do what you want, it's a free country.
```
```
Person1's intent:
Person1 expresses frustration and anger towards Person2, indicating that Person2's past actions have
caused damage to their relationship, and simply declaring love is not enough to mend it.
```
```
Person2's intent:
Person2 expresses a desire to reconnect and reestablish a father-daughter relationship with Person1.
```

Figure 4: Examples of intent summarization, illustrating how dialogue between two individuals can be analyzed to extract the underlying intent behind their statements. Each example provides a clear one-sentence summary for both Person1 and Person2, showcasing differing perspectives and emotional undertones within the conversations.