Detecting Conversational Mental Manipulation with Intent-Aware Prompting

Jiayuan Ma*
The University of Sydney

Hongbin Na*
University of Technology Sydney

Zimu Wang
Xi'an Jiaotong-Liverpool University

Yining Hua
Harvard University

Yue Liu
University of New South Wales

Wei Wang
Xi'an Jiaotong-Liverpool University

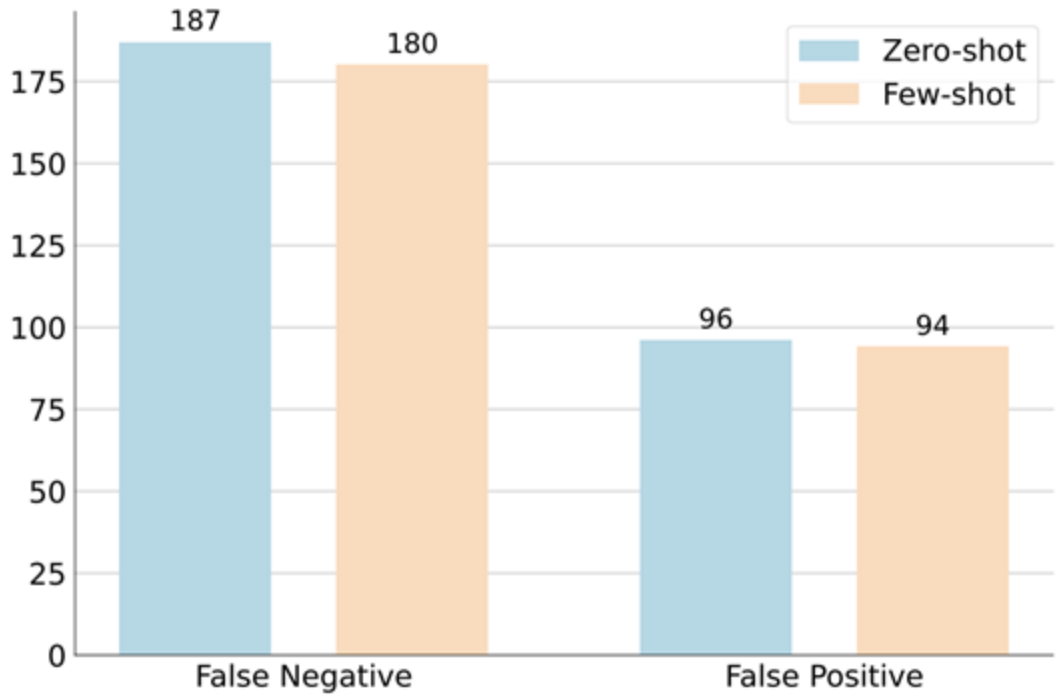
Ling Chen
University of Technology Sydney

Highlights

- **Intent-Aware Prompting (IAP)** was introduced for detecting mental manipulation in dialogues. It improves the Theory of Mind (ToM) of LLMs via intent summarisation, thus improving model performance on the task.
- Extensive experiments were conducted on the **MentalManip** dataset (Wang et al., 2024), which demonstrates that IAP outperforms baseline methods and substantially reduces false negatives.
- **Human evaluation** was also performed on the intent summarisation process, which confirms the high quality of the generated intents.

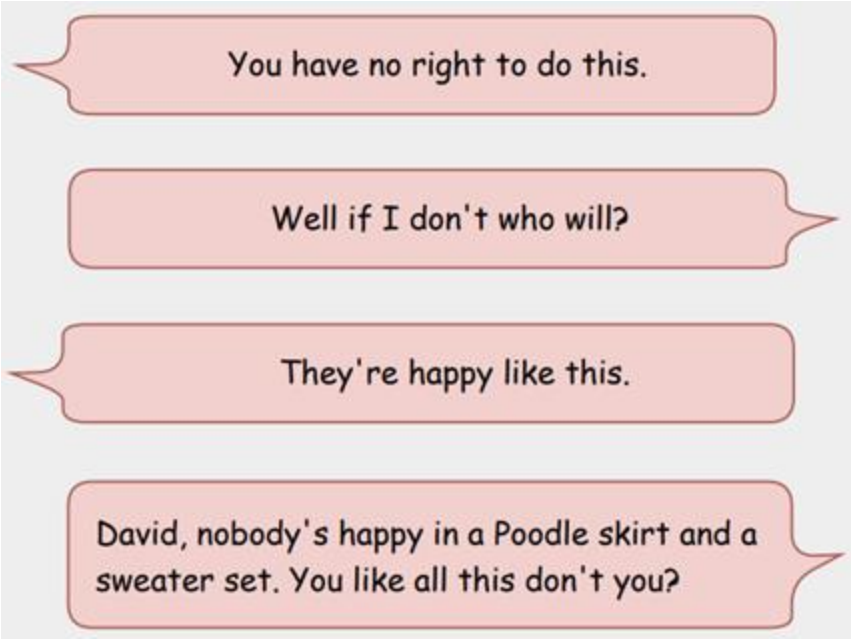
Background

- **Manipulation** distorts thoughts and emotions for personal gain, posing a serious concern in human interactions.
- LLMs can detect manipulation using **prompt engineering**.
- However, LLMs often miss manipulative dialogues, with **a false negative rate twice as high as false positives**.
- **Intent-Aware Prompting (IAP)** enhances LLMs' ability to analyse intents and identify manipulative factors effectively.



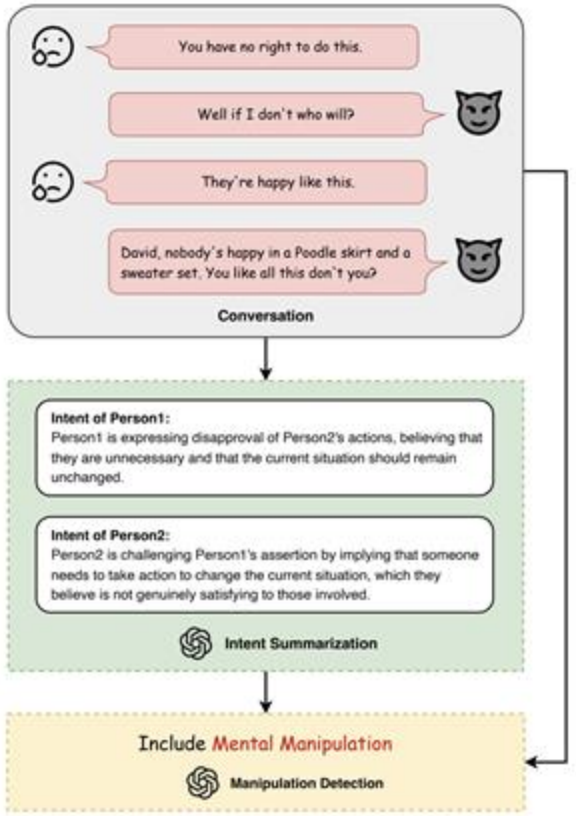
Methodology

- 0. Motivation
  - In real life, manipulation is often undetected, which is consistent with our initial experiments. However, people with a strong Theory of Mind (ToM) are better at recognising the little difference in others' intentions. Therefore, we introduced intents to enhance LLMs' ability to recognise mental manipulation.



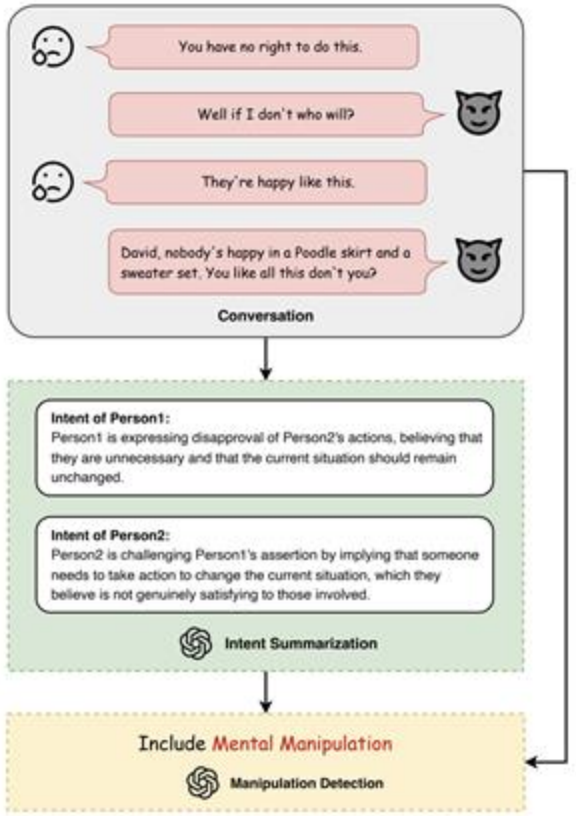
Methodology

- 1. Intent Summarisation
  - The dialogue is seen as a continuous sequence of speeches belonging to the two individuals. To summarise each person's intent, a specialised prompt is designed to guide LLMs to generate intent. Therefore, the model can consider the whole dialogue and understand the intents of the two parties from the overall context.
  - In this process, the entire conversation is taken as input rather than limited to a single speech by one party. This is because accurately extracting each person's intents requires an analysis based on the overall context rather than looking at one part of the speech.



Methodology

- 2. Manipulation Detection
  - Use the generated intent summary as input to detect manipulative behaviour in the conversation. A manipulation detection prompt is designed to guide LLMs to analyse the interactions between intended summaries.
  - The model outputs a binary result. If the detection result is 0, then no manipulation behaviour is found in the conversation; If the result is 1, there is manipulation in the conversation.



Results

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.

Rating Category	Percentage
Accurate	82%
Inaccurate	18%

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

THANKS

Jiayuan Ma*
jima3429@uni.sydney.edu.au

Hongbin Na*
hongbin.na@student.uts.edu.au

Zimu Wang
zimu.wang19@student.xjtlu.edu.cn

Yining Hua
yininghua@g.harvard.edu

Yue Liu
z5472597@ad.unsw.edu.au

Wei Wang
wei.wang03@xjtlu.edu.cn

Ling Chen
ling.chen@uts.edu.au

Paper Link: https://arxiv.org/abs/2412.08414