CultureManip: A Novel Benchmark for Mental Manipulation Detection Across Multilingual Settings

Anonymous submission

Abstract

Detecting mental manipulation is a culturally dependent and highly subjective task. We introduce CultureManip, a multilingual benchmark for manipulation detection across English, Spanish, Chinese, and Tagalog. Using raw inter-annotator agreement as our evaluation metric, we compare human–human consistency with human–LLM agreement to assess how well ChatGPT-3.5 Turbo aligns with native speakers. Human–LLM agreement is 48% in English, 41% in Spanish, 28% in Chinese, and 20% in Tagalog, revealing sharp performance drops outside of English. These results demonstrate a clear correlation between language resource availability and detection accuracy. Notably, Spanish exhibits the largest decline relative to its high human–human agreement, indicating a mismatch between model assumptions and Spanish pragmatic norms. Chinese and Tagalog show both lower human–human consistency and additional model degradation, reflecting challenges tied to indirectness, politeness strategies, and translation artifacts. These findings highlight significant cultural and linguistic gaps in current LLMs and underscore the need for culturally-aware, multilingual approaches to manipulation detection.

Introduction

Mental manipulation is the intentional use of language to influence or control someone's thoughts, emotions, or decisions (Barnhill 2014). These manipulative techniques can be subtle and difficult to detect, making them a significant challenge for both humans and artificial intelligence (AI) systems (Wang et al. 2024a; Ienca 2023). As large language models (LLMs) become increasingly integrated into digital communication, ensuring that these models can recognize and mitigate manipulative language is crucial for preventing misinformation, exploitation, and unethical persuasion As discussed by (Liu et al. 2025).

Recent efforts, such as MentalManip (Wang et al. 2024b), have primarily focused on detecting manipulative intent within English-language conversations (Ma et al. 2024; Yang et al. 2024). Studies evaluating models like GPT-4 Turbo, LLaMA-2-13B, and RoBERTa-base have revealed that state-of-the-art LLMs struggle to reliably identify manipulative content due to the inherently subjective nature of annotation and the complexity of linguistic manipulation (Wang et al. 2024a).

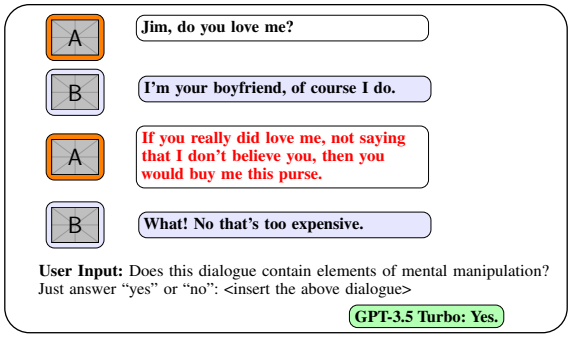

Figure 1: Example of conversation analysis by ChatGPT 3.5-Turbo.

Our research seeks to address this gap by introducing CultureManip, a multilingual benchmark designed to explore how mental manipulation manifests across different languages. Unlike previous studies that focus solely on English, our work investigates the performance of LLMs in detecting manipulation across linguistic structures and cultural contexts.

To achieve this, we conduct experiments using ChatGPT-3.5 Turbo (Brown et al. 2020) , evaluating its performance under zero-shot prompting. By examining the cross-linguistic aspects of mental manipulation, our research contributes to the development of more robust AI moderation systems, ultimately fostering safer and more ethical AI-assisted communication across a multilingual spectrum.

Related Work

Recent efforts to detect mental manipulation in language have focused heavily on evaluating and improving large language models (LLMs) using the MentalManip dataset Wang et al. (2024b). This dataset was used to conducted baseline evaluations using models like GPT-4 Turbo and LLaMA-2-13B. Their findings revealed that current state-of-the-art models struggle with reliably identifying manipulative content. The annotation process, being inherently subjective, introduces potential inconsistencies that may not align with broader societal perceptions of verbal manipulation. Furthermore, (Xiong, Gao, and Jeong 2025) demonstrated the difficulty for LLMs to classify fine-grained sarcasm, highlighting the shortcom-

ing of LLMs in detecting human conversational nuance.

Extended work on this topic examined the performance of LLMs using prompting strategies, particularly Chain-of-Thought (CoT) prompting (Yang et al. 2024). Their results demonstrated that CoT without example-based learning (i.e., zero-shot CoT) underperformed relative to simpler prompting strategies as model complexity increased.

More recently, a new novel prompting approach was proposed: Intent-Aware Prompting (IAP) (Ma et al. 2024). This prompting technique is aimed at improving the detection of mental manipulation by emphasizing the intent behind statements. IAP provided better performance over existing prompting techniques, notably through a substantial reduction in false negatives. The authors recommend the development of a more diverse and representative dataset to improve the generalization of future models.

Methods

Task Definition

In this work, we define mental manipulation as language intended to influence another person's emotions, choices, or behavior through pressure, guilt, or strategic reframing. Unlike toxicity, which often has explicit lexical markers, manipulation depends on subtle interpersonal cues and inferred intent. Annotators labeled manipulation only when an utterance attempted to shift another person's emotional or behavioral state. The same utterance may appear manipulative in one cultural setting but neutral in another, and many strategies (e.g., guilt-tripping or feigned innocence) lack clear surface markers.

Taxonomy

To construct our dataset, CultureManip, we based off of MentalManip (Wang et al. 2024b). In which our data is separated into 5 columns representing:

- ID: the identification number for each conversation block
- Conversation: Each conversation block
- Presence of Manipulation: Decide whether manipulative or not. (1 for manipulative, 0 for non-manipulative) Represents the Presence of Manipulation line in Figure 2.
- Manipulation Dialogue: whichever line(s) that specify manipulation
- Manipulation Technique: Decide what type of manipulation(s) are present based on the CultureManip taxonomy sheet (Figure 2) (Wang et al. 2024b).

Our taxonomy sheet is a replica of the taxonomy sheet shown in MentalManip however, we did not implement the Targeted Vulnerability aspect. To ensure clarity, we used the definitions given through the MentalManip paper for each manipulation technique (Wang et al. 2024b).

Data Source and Preprocessing

To support our multilingual analysis, we utilized dialogue data from the OPUS corpus (Tiedemann 2009), an open collection of parallel corpora available at https://opus.nlpl.eu.

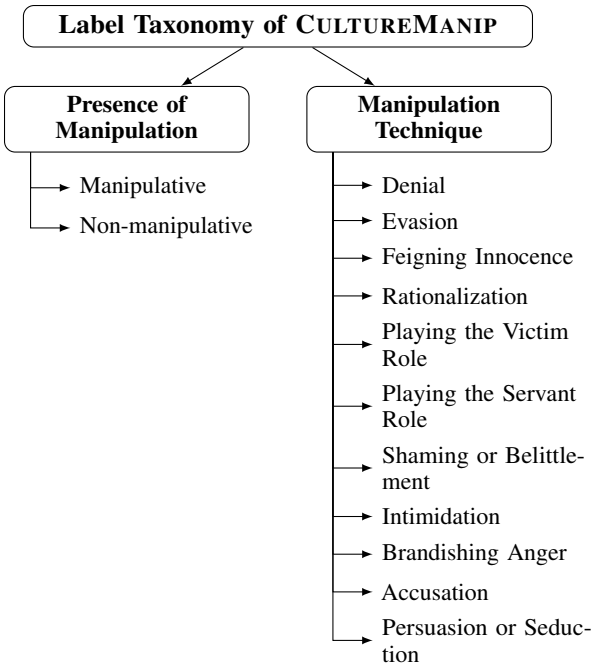

Figure 2: Multi-level taxonomy of CULTUREMANIP

Specifically, we extracted data in four languages: English, Spanish, Chinese, and Tagalog.

We used ChatGPT-3.5 Turbo to organize raw dialogue data into conversation blocks by grouping contiguous lines from the same exchange. Each block, representing a short conversation with two or more participants, was then formatted into rows in a spreadsheet.

Human Annotation

To identify instances of mental manipulation within our multilingual dataset, we enlisted two native-speaking annotators for each target language: English, Spanish, Chinese, and Tagalog. Each annotator was provided with a structured spreadsheet with 100 conversations.

The annotation protocol was standardized across languages, with annotators following shared instructions. For each conversation in Column B, annotators identified manipulative lines, transcribing them into Column C. Column D indicated manipulation presence (1 for manipulative, 0 for non-manipulative), and Column E contained selected manipulation techniques from a predefined taxonomy, including labels like Intimidation, Brandishing, and Anger. An example annotation is shown below:

- **Conversation Block:** "You're impossible, you would even delude a saint! I will divorce you!"
- **Manipulative Dialogue:** "You're impossible, you would even delude a saint! I will divorce you!"
- **Presence of Manipulation:** 1
- **Manipulation Techniques:** Intimidation, Brandishing, Anger

Disagreements were not adjudicated; instead, both labels were retained and used to compute agreement scores. Because annotators followed uniform guidelines but brought different cultural expectations into their interpretations, disagreement levels directly reflect the intrinsic variability of the task. This supports our central claim that manipulation is not uniformly perceived across linguistic and cultural contexts.

Experimental Setup

We conducted experiments in English, Spanish, Chinese, and Tagalog to measure the ability of LLMs to detect manipulation on a multilingual level. Like the human annotators, the LLM was tasked to give a binary response for each conversation, indicating whether there was presence of manipulation. All baseline experiments were carried out using zero-shot prompting, asking the model to read the conversation and detect whether manipulation is present or not, with temperature and top P set at 0.3 , as we found that these two values produced the most consistent answers.

Agreement Computation

Because manipulation is subjective, our dataset contains two independent annotations for each conversation. The annotators often disagreed, so there is no single correct label for the model to match. Raw percentage agreement is therefore the most direct measure of consistency between annotators and between annotators and the model.

For any two label sequences x and y, agreement is the proportion of conversations where the labels match:

$$\text{Agreement}(x, y) = \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}[x_i = y_i],$$

where N is the number of conversations and $\mathbf{1}[\cdot]$ is the indicator function.

Human–human agreement is computed as $\text{Agreement}(y_A, y_B)$, where y_A and y_B are the two annotators. This value reflects how often annotators agree with one another and serves as the ceiling for the task.

To evaluate the model, we compare its predictions with each annotator separately and then take the average. This avoids privileging either annotator. The resulting human–LLM agreement score is

$$\tfrac{1}{2}\Big(\text{Agreement}(y_M, y_A) + \text{Agreement}(y_M, y_B)\Big),$$

where y_M is the model prediction. The gap between human–human and human–LLM agreement reflects how closely the model approximates human judgment in each language.

Results

Multilingual Analysis

Table 1 reports agreement between human annotators (Group 1) and between the model and each annotator (Group 2) for all four languages. Because human–human agreement represents the ceiling, the key quantity is the drop from Group 1 to Group 2 rather than the absolute values alone.

Language	Group 1	Group 2
English	62%	48%
Spanish	76%	41%
Tagalog	37%	20%
Chinese	50%	28%

Table 1: Inter-annotator agreement scores for different languages. Group 1 is between two human annotators. Group 2 is the average agreement between ChatGPT-3.5 Turbo and each annotator.

Our language selection, which includes three major languages (English, Spanish, and Chinese) alongside a lower-resource language (Tagalog), was designed to highlight cross-lingual variability in both human interpretation and model alignment. Prior multilingual work such as the MEGAVERSE benchmark (Ahuja et al. 2024) and StingrayBench (Cahyawijaya et al. 2024) shows that large language models tend to perform better on high-resource languages for tasks such as sentiment analysis and named entity recognition. At first glance, our human–human agreement results seem to follow this trend, with English and Spanish showing higher consistency (62% and 76%) than Chinese and Tagalog (50% and 37%).

However, the Spanish results reveal a more interesting pattern. Despite having the highest human–human agreement, Spanish also exhibits the largest drop between human–human and human–LLM agreement (76% to 41%). This indicates that while Spanish annotators shared a stable interpretation of manipulation, the model failed to approximate that interpretation. This is likely due to pragmatic features of Spanish dialogue that the model misreads, such as emotionally expressive phrasing and rhetorical intensifiers that convey emphasis rather than coercive intent. As a result, Spanish exposes a deeper misalignment between model assumptions and culturally grounded interpretations of manipulation.

Chinese and Tagalog show lower human–human agreement, suggesting greater inherent ambiguity in how manipulation is expressed and perceived in these languages. Annotator testimonies support this, noting difficulties from indirect phrasing, translation artifacts, and context-dependent cues. The lower ceiling in these languages makes it harder for the model to achieve high agreement, yet the additional drops to 28% in Chinese and 20% in Tagalog still reflect the model's challenge in reasoning about pragmatic intent beyond surface-level features.

Spanish

Spanish communication often relies on shared context and expressive phrasing rather than explicit coercion (Hall 1976). Features such as figurative language, exclamations, and discourse markers can convey emphasis without manipulative intent (Avila and Gomez 2023). GPT-3.5 Turbo frequently misclassifies these cues, explaining the large drop between

human and model agreement.

Tagalog

Manipulation in Tagalog is often conveyed indirectly due to cultural values such as *hiya* (shame) and social harmony (Gonzalez 2010; Tupas 2015; Salazar 2011). Combined with translation artifacts in subtitle data, this indirectness reduces both human–human and human–LLM agreement.

Chinese

Chinese communication emphasizes maintaining face and avoiding direct confrontation (Ting-Toomey and Kurogi 1998; Fang and Faure 2010; Goffman 1967). Manipulation can be encoded through subtle pragmatic cues, including particles and implied obligations. Models trained largely on English-centric data struggle to capture these patterns (Hada et al. 2024), contributing to the observed drops in agreement.

Annotation Analysis

Annotator testimonies (Appendix A) show that judgments of manipulation are themselves culturally shaped, leading to label drift across languages and annotators. Several annotators noted that harsh language does not always imply manipulation, especially in Spanish where emotional expressiveness is common. In contrast, Chinese annotators often labeled subtle emotional pressure as manipulative even when English annotators did not, reflecting differences in how relational obligations and face-threatening acts are interpreted.

Tagalog annotators explicitly described translation artifacts, indirect phrasing, and context-dependent politeness strategies that affected their labeling, sometimes causing them to overlook manipulation that was present. English annotators tended to rely on explicit coercion or overt pressure, whereas Spanish annotators relied more on emotional cues and Chinese annotators relied more on implied obligations. These patterns confirm that even trained annotators do not share a universal definition of manipulation, and that some of the divergence between human–LLM and human–human agreement reflects fundamental cross-cultural subjectivity in the concept.

Manipulation Category Observations

Although the benchmark includes multiple manipulation techniques (e.g., Intimidation, Accusation, Rationalization), annotators reported that some categories were more culturally salient than others. Spanish and English dialogues more frequently exhibited overt belittlement or blame-shifting, while Chinese and Tagalog conversations displayed subtle rationalization or emotional appeals.

Implications

The disparities identified in our analysis have direct implications for multilingual safety systems. If a model systematically misses manipulation in Tagalog, Filipino users may receive weaker protection from harmful conversational patterns. Conversely, models that over-flag Spanish or Chinese speakers due to misreading expressive or idiomatic language risk subjecting these communities to disproportionate moderation or false accusations.

Such asymmetries create a form of language-driven safety inequality, where access to reliable AI moderation depends on the linguistic and cultural alignment of the user's language with the model's training distribution. In real deployment contexts, these disparities can shape whether harm is detected, ignored, or wrongly attributed. This highlights the need for manipulation-detection systems that are culturally calibrated and grounded in cross-lingual pragmatics, not merely lexical or sentiment cues.

Future Work and Conclusion

This study presents CultureManip, a multilingual benchmark for detecting verbal manipulation across languages. Using annotated conversational data, we evaluated GPT-3.5 Turbo and found large cross-lingual gaps in detection accuracy, reflecting cultural variation and the subjective nature of manipulation. These results align with prior work and illustrate the need for culturally adaptive models, since LLMs trained primarily on English perform less reliably in other languages.

Future work should refine multilingual datasets, develop clearer annotation guidelines, and study cross-lingual generalization to better understand shared and language-specific manipulation cues (Ma et al. 2025). Incorporating multimodal information such as tone or gesture may improve contextual reasoning (Herring 2015). More natural conversational sources, such as podcasts or real dialogues, could also reduce the limitations of subtitle-based data.

Although manipulation detection remains difficult for both humans and LLMs, CultureManip provides a foundation for advancing this task. By releasing the benchmark, we aim to support research toward more accurate, context-aware, and culturally sensitive models capable of handling manipulation across diverse languages.

Limitations

CultureManip has several limitations. First, data annotation is subjective, with cultural perception further complicating accurate manipulation identification. Although annotators were trained for objectivity and required to specify manipulation types, limited fluency in certain languages sometimes led to misclassifications.

Second, data sources are drawn from movie scripts, which are often stylized and exaggerated, limiting their representativeness of real-life conversations. While covering various genres, the benchmark may still reflect biases in script conventions that don't capture the full range of manipulative speech in everyday contexts.

Lastly, cultural context influences how manipulation is perceived. Different cultural norms affect how manipulative behavior is categorized, leading to potential subjectivity in labeling, as certain behaviors seen as manipulative in one culture might be overlooked or misunderstood in another.

References

Ahuja, S.; Aggarwal, D.; Gumma, V.; Watts, I.; Sathe, A.; Ochieng, M.; Hada, R.; Jain, P.; Ahmed, M.; Bali, K.; and Sitaram, S. 2024. MEGAVERSE: Benchmarking Large Language Models Across Languages, Modalities, Models and Tasks. In Duh, K.; Gomez, H.; and Bethard, S., eds., *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, 2598–2637. Mexico City, Mexico: Association for Computational Linguistics.

Avila, M.; and Gomez, J. 2023. Cultural Influences in Natural Language Processing: Challenges and Opportunities. *Journal of Multilingual and Multicultural Development*, 44(3): 245–260.

Barnhill, A. 2014. What is manipulation. *Manipulation: Theory and practice*, 50: 72.

Brown, T. B.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.; Dhariwal, P.; ...; and Amodei, D. 2020. Language Models are Few-Shot Learners. *arXiv preprint arXiv:2005.14165*.

Cahyawijaya, S.; Zhang, R.; Lovenia, H.; Cruz, J. C. B.; Gilbert, E.; Nomoto, H.; and Aji, A. F. 2024. Thank You, Stingray: Multilingual Large Language Models Can Not (Yet) Disambiguate Cross-Lingual Word Sense. arXiv:2410.21573.

Fang, T.; and Faure, G. O. 2010. Chinese communication characteristics: A yin yang perspective. *International Business Review*, 19(1): 32–42.

Goffman, E. 1967. *Interaction ritual: Essays on face-to-face behavior*. Anchor Books.

Gonzalez, A. 2010. *Language and Culture in the Philippines*. Manila: Asian Cultural Press.

Hada, R.; Gumma, V.; de Wynter, A.; Diddee, H.; Ahmed, M.; Choudhury, M.; Bali, K.; and Sitaram, S. 2024. Are Large Language Model-based Evaluators the Solution to Scaling Up Multilingual Evaluation? In Graham, Y.; and Purver, M., eds., *Findings of the Association for Computational Linguistics: EACL 2024*, 1051–1070. St. Julian's, Malta: Association for Computational Linguistics.

Hall, E. T. 1976. *Beyond Culture*. Anchor Books.

Herring, S. C. 2015. New frontiers in interactive multimodal communication. In *The Routledge handbook of language and digital communication*, 398–402. Routledge.

Ienca, M. 2023. On Artificial Intelligence and Manipulation. *Topoi*, 42(3): 833–842.

Liu, J.; Zhang, Y.; Wu, W.; Xu, W.; Xu, M.; Yang, N.; Zhou, B.; Duan, N.; Zhao, D.; Wu, Y.; et al. 2025. LLM Can be a Dangerous Persuader: Empirical Study of Persuasion Safety in Large Language Models. *arXiv preprint arXiv:2504.10430*.

Ma, J.; Na, H.; Wang, Z.; Hua, Y.; Liu, Y.; Wang, W.; and Chen, L. 2024. Detecting Conversational Mental Manipulation with Intent-Aware Prompting. arXiv:2412.08414.

Ma, W.; Zhang, H.; Yang, I.; Ji, S.; Chen, J.; Hashemi, F.; Mohole, S.; Gearey, E.; Macy, M.; Hassanpour, S.; et al. 2025. Communication Makes Perfect: Persuasion Dataset Construction via Multi-LLM Communication. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, 4017–4045.

Salazar, Z. A. 2011. Pakikisama and Social Harmony in Filipino Culture. *Philippine Journal of Psychology*, 44(1): 45–67.

Tiedemann, J. 2009. Finding Alternative Translations in a Large Corpus of Movie Subtitles. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2009)*, 449–457. Association for Computational Linguistics.

Ting-Toomey, S.; and Kurogi, A. 1998. Facework competence in interpersonal conflict: A cross-cultural comparison of Chinese, Japanese, and American cultures. *Communication Research*, 25(5): 567–590.

Tupas, R. F. 2015. Hiya and Cultural Control in the Philippines. *Asian Journal of Social Science*, 43(4-5): 490–512.

Wang, Y.; Yang, I.; Hassanpour, S.; and Vosoughi, S. 2024a. MentalManip: A dataset for fine-grained analysis of mental manipulation in conversations. *arXiv preprint arXiv:2405.16584*.

Wang, Y.; Yang, I.; Hassanpour, S.; and Vosoughi, S. 2024b. MentalManip: A Dataset For Fine-grained Analysis of Mental Manipulation in Conversations. In Ku, L.-W.; Martins, A.; and Srikumar, V., eds., *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 3747–3764. Bangkok, Thailand: Association for Computational Linguistics.

Xiong, L.; Gao, R.; and Jeong, A. 2025. Sarc7: Evaluating Sarcasm Detection and Generation with Seven Types and Emotion-Informed Techniques. In Zhang, C.; Allaway, E.; Shen, H.; Miculicich, L.; Li, Y.; M'hamdi, M.; Limkonchotiwat, P.; Bai, R. H.; T.y.s.s., S.; Han, S. S.; Thapa, S.; and Rim, W. B., eds., *Proceedings of the 9th Widening NLP Workshop*, 157–166. Suzhou, China: Association for Computational Linguistics. ISBN 979-8-89176-351-7.

Yang, I.; Guo, X.; Xie, S.; and Vosoughi, S. 2024. Enhanced Detection of Conversational Mental Manipulation Through Advanced Prompting Techniques. arXiv:2408.07676.

Annotator Testimonies

Tagalog Annotator A: " I didn't find a lot of manipulative stuff in the dialogue so far, but maybe I'm too selective? I think in the context that some are used, it's not manipulation. But I checked the Annotator B file and they were picking lines that sometimes might be manipulative, but not really."

Tagalog Annotator B: "It was not easy because the Tagalog was not translated well. Also, the reader eventually figured out what the story was. I am not sure if that's intentional, but having some idea of the story influenced the identification process of manipulation. The reader may be prone to reasoning why such statements were made and make their own justifications (unintentional but natural bias)."

Spanish Annotator A: Reading from the dialogues, I could easily tell that they were manipulative. What gave it away was the fact that they included tons of name-calling and belittlement. Checking through these conversations, I can tell the mood and tone of the dialogues. They contained exclamation points, sarcasm, and irony. This is where I figured the conversations were not nice.

Spanish Annotator B: I think it was easy to tell that most of the text was manipulative. It fit a lot of the categories given, and there was a lot of harsh language being used, which made it obvious to me that it was a manipulation.

English Annotator A: I personally found it pretty hard. It was extremely sutble for some conversation, to the point where I can make an argument for both and convince myself. The manipulation wasn't the kind that was used to control people and insult them. It was more subtle than that, so I think that ChatGPT wouldn't, or rather couldn't, do that well, but I don't know.

English Annotator B: I think most of the text was easy to tell, especially in their specific contexts. Although it might have been unclear exactly what was happening, there was a lot of chaos and dialogue that had a strong sense of manipulation that were fitting of the categories given through the taxonomy sheet.

Chinese Annotator A: I think a lot of the conversations provided had very subtle forms of Mental Manipulation. I lot of the lines weren't using mental manipulation specifically to take advantage of another person, but would fit into a lot of the categories given through the taxonomy sheet. In that sense, I believe it was a bit harder to see manipulation, and what people perceive as mental manipulation is more subjective.

Chinese Annotator B: It was a mix of good and bad. Yeah there were some conversation that were obviously manipulative, but also a good amount of conversations that could be argued for not manipulative. Overall, pretty subjective.

Ethics Statement

The benchmark was created from publicly available movie scripts, focusing on common communication forms like persuasion and emotional manipulation. Human annotators from diverse linguistic backgrounds voluntarily labeled the dialogues for manipulation, following ethical guidelines to ensure anonymity and fairness. Annotators were trained to identify linguistic cues without assuming characters' intent, considering cultural and contextual nuances. While the benchmark covers various genres and settings, it may not fully represent all cultural contexts, and future work aims to expand its diversity. The project emphasizes responsible research and ethical use, cautioning about the sensitive implications of automating manipulation detection.

Prompting

Zero-Shot Prompt

```
Zero-shot prompting format:
'''
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>
'''
```