

CHAPTER 2

LITERATURE REVIEW

2.1 Previous Studies

Related to the first question, how accurate AI agents are in identifying and categorizing social deixis based on Levinson's (1983) classification in Pramoedya Ananta Toer's *Bumi Manusia*, recent studies have examined how AIs perform in understanding linguistic pragmatics which concerns meaning in context and implied intention such as studies by Ma et al. (2025), Bojić et al. (2025), and Ruis et al. (2023).

Ma et al. (2025) provide a systematic review of research on pragmatics in LLMs. They propose a structured framework for evaluating pragmatic phenomena, including implicature, speech acts, and social norms. Ma et al. (2025) explain that AI agents are very good at producing clear and coherent text. However, their ability to understand complex pragmatic meaning such as sarcasm or implication is still limited and not well tested. This shows that we still do not fully understand how these models handle pragmatic meaning in other languages and cultural contexts. Nevertheless, as a review study, Ma et al. (2025) mainly summarizes findings from previous research and does not directly test how AI performs on specific pragmatic features. Therefore, more empirical studies are needed to examine how well AI agents can handle particular aspects of pragmatics, such as social deixis.

Next, Bojić et al. (2025) investigate how well LLMs understand conversational implicature compared to human participants. Based on Grice's cooperative principle, the researchers tested 155 participants, including both

humans and various AI models, in identifying violations of conversational maxims. Their findings show that advanced models such as GPT-4 outperformed human participants, achieving an average score of 4.8 out of 5. This study demonstrates that LLMs have developed strong capabilities in recognizing implied meaning and social cues. Although these findings show that advanced AIs can understand implied meanings, implicature and social deixis are not the same. Implicature requires understanding meanings that are implied in a conversation, while social deixis requires recognizing social relationships expressed through language. Therefore, good performance in understanding implicature does not automatically mean that AI can accurately identify social deixis.

Similarly, Ruis et al. (2023) created a task to test implicature and used an evaluation method to compare humans and LLMs. The results show that this task is still difficult for current models. The study also finds that example-level instruction tuning is a promising way to improve pragmatic understanding in LLMs. They compare different levels of model training. The results show that making the model larger does not always improve pragmatic reasoning. Instead, instruction tuning and structured reasoning methods can greatly improve performance. GPT-4 reached human-level accuracy (86.5%) when it used Chain-of-Thought prompting. The findings of Ruis et al. (2023) also suggest that AI performance in pragmatic tasks may depend not only on model capability but also on evaluation design and prompting strategies. This raises questions about whether reported improvements reflect genuine pragmatic competence or the effectiveness of task-specific prompting techniques.

Taken together, these studies suggest that AIs have shown promising performance in several pragmatic tasks, particularly implicature recognition and conversational inference. However, findings remain inconsistent. While Bojić et al. (2025) reported that GPT-4 outperformed human participants, Ruis et al. (2023) found that pragmatic reasoning remained challenging for current models and often depended on prompting strategies such as Chain-of-Thought reasoning. These mixed findings indicate that AI performance may vary across different pragmatic phenomena and evaluation settings. In addition, the differences in findings across these studies may be caused by differences in their research methods, datasets, and evaluation standards. Therefore, comparisons of AI performance between studies should be made with caution.

Although research on AI and pragmatics has increased, studies that specifically focus on social deixis are still limited. Most previous studies examine general pragmatic abilities rather than how AI identifies linguistic forms that express social relationships. Since social deixis is a pragmatic phenomenon that relies on the interpretation of social relationships and contextual information, findings from these studies provide a useful foundation for examining AI performance in social deixis classification.

These studies provide valuable insights into AI performance in pragmatic tasks. However, they primarily focus on implicature and conversational inference rather than social deixis. Consequently, little is known about how accurately AI agents identify and categorize social deixis based on established pragmatic frameworks such as Levinson's (1983) classification, particularly in literary texts.

Therefore, the present study extends previous research by evaluating how accurately AI agents identify and categorize social deixis based on Levinson's (1983) classification. By focusing on a specific pragmatic category, this research contributes to a more detailed understanding of AI performance in context-sensitive linguistic analysis.

Additionally, related to the second research question, how accurate are AI agents in translating social deixis based on their selection of Molina & Albir's (2002) Translation Techniques in Pramoedya Ananta Toer's *Bumi Manusia*, several previous studies have examined the translation of deixis and address terms in literary texts by focusing on translation techniques used by human translators such as Riyananda (2017), Hasanah (2018), and Isdhianty (2021).

Riyananda (2017) analyzed address terms using Guerra's framework and found six techniques and the frequently used techniques were borrowing and literal translation. Several of these techniques, such as borrowing, literal translation, modulation, and transposition, are also included in Molina & Albir's (2002) model. This shows a methodological similarity with the present study in terms of technique types. Hasanah (2018) studied all types of deixis translation and identified nine techniques by Molina & Albir's (2002) framework, with literal translation being the dominant technique. Similarly, Isdhianty (2021) examined the translation of address terms in a novel. The study found that the translator used single techniques and combinations of two techniques by Molina & Albir's (2002) scheme. The frequently used techniques are established equivalent and pure borrowing.

Substantially, these studies demonstrate that the translation of deixis and address terms has commonly been analyzed through translation techniques, particularly those proposed by Molina and Albir (2002). The findings consistently show the frequent use of techniques such as literal translation, borrowing, and established equivalent. However, these studies focus exclusively on human translators and do not examine how AI agents select translation techniques when translating socially marked expressions.

Another common finding in these studies is that translators often use literal translation and borrowing when translating deixis and terms of address. This shows that they try to keep the original social meaning while making the text understandable in the target language. However, how well these techniques work may depend on the cultural and linguistic differences between the source and target languages.

To address this limitation, several studies have compared the performance of human translators and AI agents in literary translation, such as those conducted by Voita et al. (2019) and Webster et al. (2020).

Voita et al. (2019) analyzed several linguistic aspects, such as deixis, ellipsis, and lexical cohesion. Their findings show that machine translations can be correct at the sentence level but become inconsistent when the sentences are read together in a wider context and fail to handle discourse-level meaning properly. In addition, Webster et al. (2020) compared human translation and neural machine translation by analyzing textual features in translated texts. Their study shows that human translations of literary texts use richer vocabulary, show better cohesion,