The Relational Prior: Unlocking Social Intelligence in Large Language Models by Reasoning About Relationships

Anonymous ACL submission

Abstract

Understanding the relationship between individuals is fundamental to social intelligence. However, language models often struggle with social N because they overlook these critical relational dynamics. This paper introduces Guided Relational Context AS Prior (GRASP) framework that enhances social understanding by explicitly instructing the model to first infer the relationship between speakers, the *relational prior*, before performing a downstream task. We evaluate this approach on two challenging tasks, mental manipulation detection and social norm violation detection, across five state-of-the-art LLMs. Our experiments demonstrate that GRASP consistently improves performance in zero-shot settings, often outperforming even few-shot baselines. This study highlights that modeling interpersonal relationships is a powerful and efficient method for unlocking more robust social intelligence in language models.

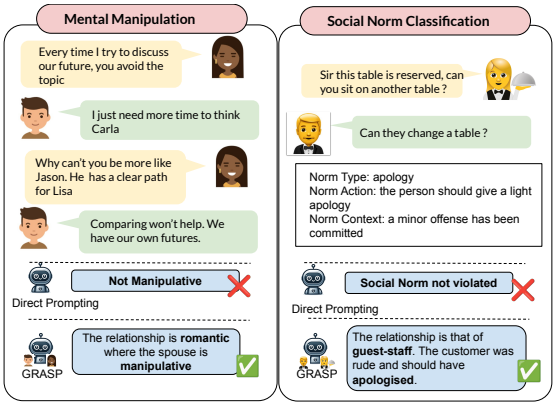


Figure 1: (Left) An example from the manipulation detection task. Recognizing the romantic relationship between speakers allows GRASP to correctly classify this conversation as manipulative. (Right) An example from the social norm classification task. Recognizing the guest–staff relationship allows GRASP to see that the interaction requires formality, which the reply lacks, leading to its correct classification as a norm violation.

1 Introduction

Human behavior varies significantly depending on the relationship between participants (Hinde, 1976, 1987), and this, in turn, shapes what we can infer from a conversation ('*social understanding*') (Frith, 2007). For example, a friend saying '*You'd go if you really cared about me*' may be seen as emotional manipulation, while the same line from a romantic partner might be interpreted more empathetically. Similarly, a casual, offhand remark may be acceptable between close friends but inappropriate in a professional context.

As these relational cues shape how humans use language, they must also inform how technologies like LLMs engage in conversation. As these technologies become more integrated into our lives, they must develop social intelligence. For this, they must understand social relationships, as these relationships dictate the unspoken norms and expectations that govern our interactions. Without this context, their responses risk being tone-deaf or offensive, diminishing their utility and trustworthiness.

Most prior work on social intelligence of LLMs has primarily focused on directly prompting models to perform the downstream task (Wang et al., 2024; Zhan et al., 2023). Some recent approaches explore improved prompting strategies such as intent-aware prompting (Ma et al., 2025b). However, these methods largely overlook the role of interpersonal relationships in social understanding, especially in dyadic conversations. In such conversations, the interaction is shaped almost entirely by the dynamics between the two participants. In such settings, the tone, intent, and meaning of an utterance hinge on who is speaking to whom. For example, the (left) conversation in Fig. 1, seems as expressing concern or constructive comparison. However, when the relationship is correctly inferred as romantic, the same expressions (e.g., guilt induction, social comparison) signal emotional pressure, indicating mental manipulation by one of the speakers.

However, modeling relationships is challenging. Relationship cues in dialogue are often subtle, implicit, and context-dependent. This is especially challenging for smaller LLMs, which struggle at understanding social interactions (Shi et al., 2023). For such LLMs, to truly understand social scenarios, they must explicitly model social relationship.

This is the key idea behind this paper: the idea that inferring relationships is important for social understanding tasks. To operationalize this idea, we propose Guided Relational Context AS Prior (GRASP)—a two-step social reasoning framework that first characterizes the relationship between the dialogue participants and then uses this as a prior for performing the downstream social understanding tasks.

We evaluate GRASP on multiple state-of-the-art LLMs from different families and across two challenging social understanding tasks: mental manipulation detection and social norm violation detection. Our experiments lead to two key observations. First, GRASP consistently outperforms other baselines, including general reasoning-based baselines. This demonstrates the importance of relationship-based reasoning. Second, GRASP, in a zero-shot setting, outperforms even few-shot baselines. This demonstrates that the relational prior is more useful than even manually labeled in-context demonstrations. We further conduct ablation studies and examine how performance varies with relationship type and dialogue length. Finally, we present an error analysis of failure cases to highlight the limitations of GRASP and directions for improvement.

Our contributions are the following.

1. We propose the idea that inferring relationships is important for social understanding in NLP technologies.
2. We present GRASP, an extensible framework that uses relational prior to improve social understanding tasks.
3. We introduce innovative prompting strategies that mitigate common LLM biases, such as agreeableness bias.
4. Our comprehensive experiments demonstrate the utility of including relational prior across diverse tasks, models, and baselines, including outperforming other reasoning baselines.

2 Related Work

Social understanding in dialogues is a long-standing challenge intersecting with areas like pragmatics (Ma et al., 2025a), discourse analysis (Joty et al., 2019), and computational social science (Hou and Huang, 2025; Dayanik, 2022). Our work is related to the following strands of research.

Social and Affective Understanding. Work on emotion recognition (Poria et al., 2019; Majumder et al., 2019), empathy modeling (Buechel et al., 2018), rationale generation (Dutt et al., 2024), and affective response generation (Rashkin et al., 2019; Zhou et al., 2018) demonstrates the importance of using context of social cues(Morrill et al., 2024) in dialogue classification. Although these approaches often use emotion or personality labels, they rarely include interpersonal relationships, which are critical for inferring subtle cues for complex social understanding tasks.

Manipulation and Persuasion Detection. A growing body of work has focused on identifying manipulative, persuasive, or deceptive language in text. Wang et al. (2023) introduces the MentalManip dataset for fine-grained manipulation detection in movie dialogues. Other studies have examined argumentative strategies in persuasion (Hidey et al., 2017), or intent(Sheshanarayana et al., 2025; Yerukola et al., 2024; Kim et al., 2024; Qian et al., 2024) in negotiation and online content (Chen et al., 2020). However, these approaches typically treat manipulation as a surface-level label, without modeling the underlying relationship dynamics that give rise to manipulative behavior.

3 Task Definition

Let a dyadic dialogue D between two speakers, s_1 and s_2, be defined as:

$$D = \{u_1, u_1, \ldots, u_n, u_n\}$$

where u_i denotes the ith utterance. Given such an input dialogue, D, our goal is output the classification label for the dialogue $c \in \{1, 2, 3...K\}$.

In this work, we illustrate our idea by experimenting with two specific social understanding tasks: mental manipulation detection and social norm classification.

Mental manipulation detection requires classifying whether the dialogue, D, involves one of the speakers mentally manipulating the other speaker. The label set is $C = \{$Manipulative, Non-Manipulative$\}$. Figure 1 (left) shows an example containing a dialogue and the corresponding label.

Social norm classification requires classifying whether the dialogue, D, violates a given social

norm. The label set is $C = \{\text{Violation, Adherence}\}$. Each example is paired with a structured norm rule comprising a *Norm Type*, *Norm Context*, and *Norm Action*. Figure 1 (right) shows an example of the task.

4 Social Reasoning with Relational Prior

In this section, we describe our central proposal, GRASP (Sec. 4.1) and its extension (Sec. 4.2).

4.1 Core Framework: GRASP

The idea behind GRASP is motivated by insights from cognitive science (Fiske and Haslam, 1996; Frith, 2007), which emphasize that understanding the nature of interpersonal relationships is central to social understanding. Similar to them, we hypothesize that LLMs benefit from explicitly modeling speaker relationships as an intermediate step for social understanding of dialogues. To evaluate this, we present, GRASP (Guided Relational Context AS Prior), a two-step decomposition strategy that establishes relational context and uses that as a prior for addressing the social understanding tasks.

Step 1: Relationship Inference. In this step, GRASP prompts the LLM to analyze the dialogue, D, and generate a natural language description of the relationship, r, between the speakers s_1 and s_2:

$$r = \text{LLM}(D, \mathcal{P}_{\text{rel}}(s_1, s_2))$$

where $\mathcal{P}_{\text{rel}}(s_1, s_2)$ is the prompt that instructs the model to infer the relationship between the speakers.

> $\mathcal{P}_{rel}$: *Describe the relationship between the speakers.*

Step 2: Contextualized Prediction. Next GRASP uses the inferred relationship r as additional prior for the final classification:

$$c = \text{LLM}(D, r, \mathcal{P}_c)$$

where $\mathcal{P}_c$ is the prompt that elicits the final classification output conditioned on the relationship. For example, $\mathcal{P}_c$ for the task of mental manipulation is:

> $\mathcal{P}_c$: *Based on the relationship, please determine if the dialogue contains elements of mental manipulation, by answering '[a]' for Yes or '[b]' for No. Your output must end with a single sentence formatted as: "Therefore, the answer is [a]" or "Therefore, the answer is [b]"*

The complete prompt for GRASP is shown in Appendix A.1. There are two key innovations offered by GRASP:

First, the final classification prompt explicitly directs the model to treat the inferred relationship as a prior for its subsequent reasoning using the phrase 'Based on this relationship...' This acts as a form of soft conditioning, where the inferred relationship serves as a latent contextual variable that guides the prediction of the final label. This setup mirrors the structure of latent variable models that have proved to be useful in complex reasoning tasks (Zhou et al., 2023; Khot et al., 2023). Such conditioning improves the model's reasoning ability, especially in cases where similar cues carry different implications depending on who the speakers are.

Second, most prior works ask the model to indicate its classification output in simple, open-ended answer formats, such as responding with a 'Yes' or 'No'(Wang et al., 2024). Such formats are unfit for social tasks as they are susceptible to the inherent biases of instruction-tuned models, like agreeableness bias (Itzhak et al., 2024), where the model may be more likely to respond affirmatively regardless of the content. We overcome this challenge by making the LLM choose from a predefined and neutral set of options '[a] or [b]'.

4.2 Multi-Step Reasoning with GRASP_Ext

The GRASP framework is extensible. More complex social reasoning chains can be constructed by adding further intermediate social reasoning steps closely related to the task at hand. We demonstrate this with an extension for the mental manipulation task: GRASP_Ext. In between the two steps of GRASP, this variant adds a step to assess the level of persuasion, a concept closely related to manipulation, conditioned on the inferred relationship.

$$r = \text{LLM}(D, \mathcal{P}_{\text{rel}}(s_1, s_2))$$
$$p = \text{LLM}(D, r, \mathcal{P}_{\text{pers}})$$
$$c = \text{LLM}(D, r, p, \mathcal{P}_c)$$

> $\mathcal{P}_{pers}$: *Describe whether the dialogue contains a typical level of persuasion that is common in everyday conversations.*

The prompt for GRASP_Ext is shown in Appendix A.2.

Experiment Setting	Llama 3.1 8B	Qwen 2.5 7B	Gemma-2 9B	Mistral 7B	Llama 3.1 70B	Overall performance
DIRECT_WANG	41.86 ± 0.46	58.51 ± 0.59	42.82 ± 0.70	43.91 ± 0.72	63.22 ± 0.65	49.48
IAP	48.65 ± 1.07	51.52 ± 0.55	49.30 ± 0.73	56.41 ± 1.24	64.05 ± 0.72	53.65
CoT	42.93 ± 0.75	59.02 ± 1.17	51.10 ± 1.46	**63.42 ± 1.41**	64.88 ± 1.10	56.27
Self-Explain	42.56 ± 0.42	60.75 ± 0.66	43.18 ± 0.61	48.51 ± 0.96	63.95 ± 0.83	51.79
Self-Ask	32.75 ± 1.14	60.19 ± 1.86	48.75 ± 1.05	50.02 ± 0.75	64.12 ± 1.21	51.17
DIRECT	51.92 ± 1.31	52.11 ± 0.95	46.03 ± 0.96	**59.40 ± 0.99**	65.01 ± 0.74	53.77
GRASP	57.22 ± 0.95	**63.74 ± 1.04**	54.34 ± 1.37	61.16 ± 0.45	**67.45 ± 0.62**	60.37
GRASP_EXT	**62.62 ± 1.32**	**64.46 ± 0.67**	**62.77 ± 0.83**	**61.29 ± 1.17**	**68.10 ± 0.55**	**63.26**

Experiment Setting	Llama 3.1 8B	Qwen 2.5 7B	Gemma-2 9B	Mistral 7B	Llama 3.1 70B	Overall performance
DIRECT	49.03 ± 2.26	50.56 ± 0.45	61.30 ± 0.20	49.65 ± 0.06	63.10 ± 0.48	54.73
CoT	**57.52 ± 2.30**	53.41 ± 1.07	60.74 ± 0.48	49.51 ± 0.70	64.55 ± 0.91	57.55
Self-Explain	56.79 ± 0.71	54.91 ± 0.26	59.45 ± 0.27	49.25 ± 0.12	63.88 ± 0.64	56.86
Self-Ask	54.74 ± 1.18	57.35 ± 1.60	59.11 ± 0.74	45.57 ± 0.50	64.00 ± 0.70	56.15
GRASP	56.54 ± 1.31	**57.38 ± 0.73**	**62.25 ± 0.45**	**55.19 ± 0.16**	**65.12 ± 0.59**	**59.70**

Table 1: Zero-shot macro F_1 scores for mental manipulation detection (top), and social norm classification (bottom). Bold indicates statistically significant improvement. GRASP and GRASP_EXT significantly outperform the baselines.

5 Experimental Setup

5.1 Datasets

MentalManip. This dataset was proposed by Wang et al. (2024) for the task of detecting mental manipulation. It consists of 2915 dialogues derived from movie scripts, featuring dyadic exchanges. Each dialogue is annotated for whether it contains elements of mental manipulation (binary annotation). We consider only those examples where all three annotators unanimously agree on the presence or absence of manipulation.

SocialDial. The SocialDial dataset (Zhan et al., 2023) consists of 1,563 human-written dialogues grounded in Chinese cultural norms, spanning five core social norm categories: *greetings, requests, apologies, persuasion*, and *criticism*. Each dialogue comes with norm type, action, and context, and is annotated with a norm violation (or adherence) tag. These tags are at utterance level but since each dialogue contains only one utterance with a violation tag, we treat this as the dialogue-level label. We use the publicly available dialogues[1], translating them to English using the `facebook/nllb-200-distilled-600M` model (Costa-jussà et al., 2022). After translation and cleaning, to remove incomplete or noisy examples, we retain 1,468 dialogues.

[1] https://github.com/zhanhl316/SocialDial

5.2 Baselines

DIRECT is our primary baseline that is equivalent to the Contextualized Prediction step of GRASP (i.e. without the relationship inference step). For mental manipulation task, we also compare with baselines specifically proposed for it - **DIRECT_WANG (Wang et al., 2024)**. and **IAP (Ma et al., 2025b)**. DIRECT_WANG asks the LLM to directly classify manipulation in a 'Yes' or 'No' format. IAP extends DIRECT_WANG by asking the LLM to infer the *intentions* of both speakers before determining whether the dialogue contains mental manipulation.

We also include reasoning-based prompting strategies that introduce an intermediate step prior to the social understanding task, but which do not explicitly model relationships. In **CoT (Wei et al., 2022)**, the model is instructed to reason 'step by step'. In **Self-Explain (Gao et al., 2023)**, the model first explains each utterance in the dialogue before predicting the label. Finally, in **Self-Ask (Press et al., 2022)**, the model generates and answers follow-up sub-questions that decompose the overall task before arriving at the final decision. Complete prompts for all baselines are in Appendix A.3.

5.3 Implementation Details

We experiment with the following LLMs: Llama 3.1-8B-Instruct (AI@Meta, 2024), Qwen 2.5-7B-Instruct (Qwen Team, 2024), Gemma-2-9B-it (DeepMind@Google, 2024), Mistral-7B-Instruct (Jiang et al., 2023), and Llama 3.1-70B-Instruct-Quantized (AI@Meta, 2024). See Sec. D for more

details regarding GPU, temperature, and run details.

6 Results

6.1 Effectiveness of Relational Prior in Zero-shot settings

We first compare GRASP and its extension with the baselines. Table 1 shows the macro F1 scores for zero-shot reasoning for the tasks of mental manipulation (top) and social norm classification (bottom), respectively. The last column represents the overall averaged scores across all models. The scores in bold indicate statistically significant improvement over the next best model (paired t-test; p-value< 0.05). Comparable performance is indicated by multiple scores in bold. We can make the following observations from these tables.

First, we observe that DIRECT is outperforming DIRECT_WANG by a large margin in 3 out of 5 LLMs. It is even outperforming IAP, which employs additional intermediate reasoning steps. This suggests that the proposed formatting strategies (providing predefined output options and formatting) are helping the model overcome the inherent agreeableness bias of instruction-tuned models.

Second, we observe GRASP outperforms DIRECT and DIRECT_WANG across almost all LLMs in both tasks. This highlights the benefit of including social relationships as a prior for tasks requiring complex social inference.

Third, to demonstrate that this improvement is not simply due to extra reasoning steps, we compare GRASP to general reasoning baselines - CoT, Self-Explain and Self-Ask, and also task-specific reasoning baseline, IAP. We see that GRASP consistently outperforms these reasoning baselines. For both tasks, GRASP improves over the four baselines with statistical significance, in all but one model.

These results show that it is not merely the addition of intermediate reasoning steps that benefits social understanding. Rather, explicitly incorporating relationship information provides *a distinct and more effective improvement* that other forms of reasoning are unable to provide.

Fourth, we observe that in Table 1, GRASP_EXT performs best among all methods. This suggests that incorporating additionally task-specific reasoning can further benefit social understanding tasks.

We also experimented with a GRASP variant that only incorporates typical persuasive behavior (and not relationships). It did not outperform GRASP_EXT. This indicates that the relational prior is contributing to the majority of performance gains. Notably, in some cases the performance of GRASP_EXT and GRASP are comparable, indicating the utility of just reasoning about relationships.

Lastly, we analyze the performance of individual LLMs. Llama-3.1 (8B) and Qwen-2.5 (7B) demonstrate the highest gains from GRASP, improving F1 by 6-8 points over DIRECT. Mistral-7B, Gemma-2 (9B), and Llama 70B also show significant improvements for one of the two tasks and moderate improvements or comparable performance for the other task. We made similar observations for GPT 4o mini (not shown here due to space constraints). This suggests that stronger models may already internalize some aspects of social relationships implicitly but not for all tasks. Hence, it is important to incorporate the relational prior in their reasoning.

Overall, these results reaffirm our central hypothesis: relational prior is a robust and transferable inductive bias that enhances LLM performance across social understanding tasks.

6.2 Comparison with Few-Shot methods

We next evaluate GRASP and its extension in a more challenging setting where we compare their zero-shot performances with few-shot performances of the baselines. The shots were randomly sampled and we provided the gold labels of the respective social understanding task in context. Table 2 presents the F1 scores.

We observe that in general, reasoning with the relational prior in zero-shot settings matches or exceeds the performance of few-shot baselines, often with significant margins. For instance, for mental manipulation (top), Llama-3.1 achieves outperforming the best 4-shot baseline by 8 points. Similarly, for social norm classification (bottom), for all models, GRASP in the zero-shot setting outperforms both 2-shot and 4-shot DIRECT by about 3-6 points.

This is a strong result: *even without access to in-context demonstrations*, explicitly guiding the model to reason about the relationship between speakers allows it to perform better than traditional in-context learning approaches.

7 Ablation Studies

We now analyze the effect of relational prior on dialogues of different relationships and lengths.

Experiment Setting	Shots	Llama 3.1 8B	Qwen 2.5 7B	Gemma-2 9B	Mistral 7B	Llama 3.1 70B	Overall Performance
DIRECT_WANG	2	48.23 ± 3.22	62.68 ± 2.57	59.48 ± 2.24	60.52 ± 2.86	68.04 ± 2.18	59.79
DIRECT	2	53.49 ± 4.25	62.87 ± 0.90	59.98 ± 2.42	61.20 ± 4.74	65.55 ± 3.87	60.62
DIRECT_WANG	4	53.21 ± 2.10	64.01 ± 1.92	60.74 ± 0.87	61.89 ± 2.36	63.45 ± 1.04	60.66
DIRECT	4	54.55 ± 3.11	62.92 ± 1.83	60.72 ± 2.52	63.17 ± 2.30	67.57 ± 2.30	61.79
GRASP	0	57.22 ± 0.95	63.74 ± 1.04	54.34 ± 1.37	61.16 ± 0.45	65.38 ± 0.59	60.37
GRASP_EXT	0	**62.62 ± 1.32**	64.46 ± 0.67	62.77 ± 0.83	61.29 ± 1.17	65.15 ± 0.67	**63.26**

Experiment Setting	Shots	Llama 3.1 8B	Qwen 2.5 7B	Gemma-2 9B	Mistral 7B	Llama 3.1 70B	Overall Performance
DIRECT	2	53.4 ± 0.66	51.42 ± 2.08	55.88 ± 3.43	50.33 ± 0.62	57.47 ± 0.78	53.70
DIRECT	4	49.77 ± 2.14	49.94 ± 1.23	54.98 ± 0.53	50.66 ± 1.66	56.77 ± 1.59	52.42
GRASP	0	**56.54 ± 1.31**	**57.38 ± 0.73**	**62.25 ± 0.45**	**55.19 ± 0.16**	56.00 ± 0.77	**57.47**

Table 2: Comparison with few-shot baselines for mental manipulation detection (top), and social norm classification (bottom). Bold values indicate statistically significant improvement. GRASP and GRASP_EXT in zero-shot settings significantly outperform the few-shot baselines or perform comparably.

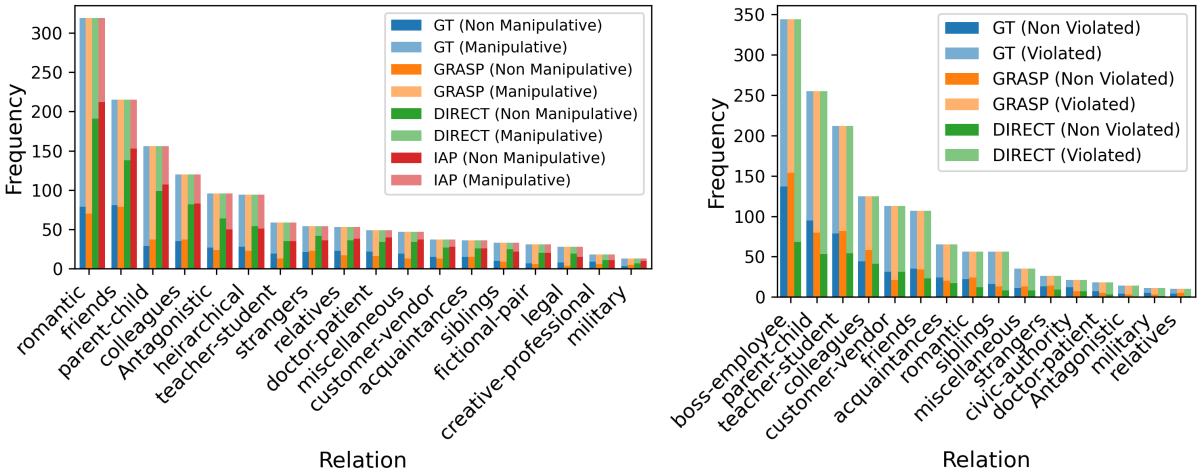


Figure 2: Label distribution for different relations for both datasets. Each cluster of bars corresponds to a specific relation, with bar colors indicating the method and class label. Color shading distinguishes between manipulative (or violated) and non-manipulative (or adhered) outcomes within each method.

7.1 Impact of Relationship Type

We study whether the effectiveness of GRASP varies across different types of social relationships. For this, we consider the inferred relationship between the speakers in each dialogue.

We conduct this analysis with Qwen-7B for the MentalManip dataset and Gemma-9B for the SocialDial dataset respectively, as these models achieved the highest F1 scores on their respective tasks. We map inferred relations to the standardized taxonomy of Zhao et al. (2024) and group dialogues accordingly to evaluate relation-specific performance.

Figure 2 displays the frequency of different relations for the MentalManip (left) and SocialDial (right) datasets. Each bar represents a relationship and is split by the predicted (or ground truth) labels. Darker shades represent the manipulative (or violated) samples. The corresponding figures for all models are shown in Appendix Fig. 5.

We make two key observations from these figures. First, the distribution of relationship types varies significantly between the two datasets. The MentalManip dataset predominantly features personal relationships, such as romantic partners, friends, and family members, while the SocialDial dataset contains more professional or formal relationships, including boss-employee and teacher-student dynamics. This highlights the diversity and domain-specific nature of relationship types across tasks, along with the robustness of GRASP.

Second, the distribution of predicted labels differs notably across methods. DIRECT and IAP overpredict the non-manipulative class (for mental manipulation) and violated class (for social norm

classification). By contrast, GRASP aligns more closely with ground-truth distributions across relations and tasks, indicating strong generalization. Appendix fig. 4 shows similar findings on manually annotated relationships for the SocialDial dataset.

We also analyzed the F1 scores for different relations. Figures 6 and 7 in the Appendix show the F1 score per relationship in the two datasets. The figures show that using the relational prior yields F1 gains for almost all relations.

To control for data imbalance, we introduce a normalized $\Delta F1$ metric that weights improvements by relation frequency (see Appendix B.1 for definition). Table 3 presents $\Delta F1$ across different relationships for both the tasks. We can see that

MentalManip		SocialDial	
Relation	**ΔF1**	**Relation**	**ΔF1**
romantic	3.43	colleagues	1.40
colleagues	1.54	teacher-student	1.34
friends	1.46	boss-employee	1.23
parent-child	1.28	customer-vendor	1.11
customer-vendor	0.77	parent-child	0.94
doctor-patient	0.50	miscellaneous	0.42
legal	0.46	romantic	0.39
hierarchical	0.45	relatives	0.24
relatives	0.40	strangers	0.23
teacher-student	0.35	antagonistic	0.10
acquaintances	0.33	friends	0.01
antagonistic	0.31	civic-authority	0.00
fictional-pair	0.31	siblings	-0.07
siblings	0.26	military	-0.10
strangers	0.19	doctor-patient	-0.12
creative-professional	0.06	acquaintances	-0.15
military	-0.05	–	–

Table 3: ΔF1 scores across different relations for Qwen

most $\Delta F1$ scores are positive, showing that relational priors help across diverse relations. In mental manipulation detection, the largest gains occur in symmetric, emotionally driven relations such as *romantic*, *friends*, and *colleagues*, where manipulative intent depends heavily on social dynamics. For example, a phrase like "You never listen to me" may be harmless teasing among friends but coercive in a romantic partnership. Inferring correct relationships is important in such cases. For social norm classification, the greatest gains appear in formal asymmetric relations such as *teacher–student* or *boss–employee*, where norms are well-defined by role expectations. Gains are smaller for ambiguous relations (e.g., *strangers*, *acquaintances*, *fictional-pair*), where norms are less clearly defined.

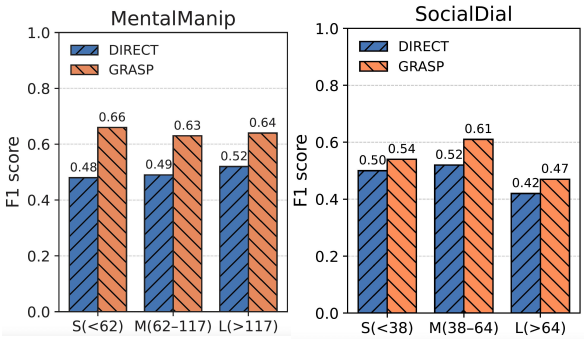


Figure 3: F1 scores of DIRECT and GRASP for dialogues of different lengths on MENTALMANIP and SOCIALDIAL. GRASP consistently improves over DIRECT, with especially large gains for shorter dialogues.

7.2 Impact of Dialogue Length

To investigate whether the effectiveness of GRASP varies with dialogue length, we grouped instances into short, medium, and long categories based on token count. Figure 3 plots the F1 scores of GRASP and DIRECT for these dialogue groups. We observe consistent improvements in both tasks from GRASP across all length buckets. However, the gains are slightly more pronounced for short and medium-length dialogues (especially those under 70 tokens) as compared to longer dialogues. This could be because while all kinds of dialogues benefit from relational prior, short and medium dialogues benefit especially more because the relational prior provides the additional context needed to make the inference. Longer dialogues might sometimes have enough information in the content of the dialogue itself to implicitly draw these inferences.

For SocialDial, we observe that the improvements are less pronounced for very short dialogues (38 tokens). We hypothesize that very short dialogues might have too few utterances, making it harder to infer relations. In contrast, longer (and medium) dialogues may offer enough context for the model to infer relations.

Relationship	Acc.	Relationship	Acc.
chief-subordinate	84	mentor-mentee	84
commander-soldier	69	partner-partner	76
customer-server	69	peer-peer	46
elder-junior	30	student-professor	69
		Overall	**66.35**

Table 4: Accuracy across different gold relationships. (GRASP w/ Gold Relations)

Model	GRASP w/ Gold Relations
Llama-3.1-7B	58.59 ± 0.44
Qwen-2.5-7B	56.41 ± 1.39
Gemma-2-9B	61.33 ± 0.84
Mistral-7B	50.19 ± 0.24
Llama-3.1-70B	66.31 ± 1.14
Overall	58.57

Table 5: F1 scores on SocialDial with gold relations.

7.3 Experiments with Gold Relationships

The SocialDial dataset comes with human annotations for speaker relationships in each conversation ("gold relationship"). We use them to study the accuracy of relationship inference. For each conversations, we compare the predicted relationship to the gold relationship using GPT-4 API. We needed to use an LLM because the taxonomies of the gold and predicted relationships are different. The accuracies are listed in Table 4. For most relationships, we observe a significantly better-than-random accuracy ($\sim$ 24.7%), indicating that the predicted relationships match human-annotated relationships.

In some practical applications, like counselling, the true speaker relationship might be known. Here, we emphasize the value of explicitly using relationship information for improved social understanding. To demonstrate this empirically, on SocialDial, we evaluate models with gold relationships provided instead of inferring them. As shown in Table 5, using gold relations consistently match or outperform Direct (e.g., Llama-3.1-7B improves from 49.0 to 58.6 F1). Interestingly, GRASP with inferred relationships remains competitive, sometimes surpassing the gold-relation version, highlighting the advantage of contextual and fine-grained cues over fixed taxonomies.

8 Qualitative Analysis

For a qualitative analysis, we examine instances where GRASP is correct but Direct is not. Qualitative examples in Appendix Table 6 are examples where GRASP predicts the final task label correctly but the Direct baseline fails. In these examples, GRASP correctly leverages inferred relationships to disambiguate tone and context-sensitive cues, such as distinguishing concern from coercion in personal settings or identifying formal role-based norms in professional settings.

We also performed an error analysis for GRASP, and found that incorrect predictions can be grouped into three categories. Examples for each category are provided in Table 7.

Category 1: Relationship inferred incorrectly. Since inferring a relationship is fundamental to social understanding, errors in inference propagate to the final task. For instance, when a peer-peer interaction is misclassified as hierarchical, GRASP incorrectly mislabels behavior that should be judged as norm violation.

Category 2: Relationship excuses behavior. Sometimes, even when the relationship is inferred correctly, GRASP misclassifies the conversation by treating it as acceptable within that relationship. For example, in certain romantic contexts, highly emotive or coercive language is misjudged as "typical for a romantic relationship." In contrast, Direct correctly predicts manipulation in these cases.

Category 3: Relationship is uncertain. In this category, both the relationship inference and the downstream social understanding task are inherently challenging and ambiguous, even for humans. These cases mostly occur in short dialogues, though not exclusively. In such settings, the inferred relationship is largely a guess due to minimal information, which limits its usefulness for the final task. Importantly, this category is also difficult for the Direct. For example, Figure 3 shows that GRASP and Direct perform comparably on short conversations

9 Conclusion and Future Directions

Our work highlights the critical role of relationship understanding in socially grounded tasks. We introduce GRASP, which leverages relational context as a prior to enhance the social reasoning capabilities of LLMs by prompting them to first infer interpersonal relationships. Our findings consistently support the hypothesis that modeling social relationships yields general-purpose gains across tasks and settings, including both zero-shot and few-shot regimes. An additional strength of our evaluation is that the Chinese origin of SocialDial provides a proxy for evaluating social understanding across cultural contexts. Future work can extend this framework by incorporating self-revision mechanisms or applying it to new domains like trust estimation, deception detection, and conflict resolution, where social dynamics play a critical role.

Limitations

While our findings are promising, this study has its limitations. Our strategies assume well-structured,

turn-by-turn dialogue with clearly defined speaker roles. In real-world deployments such as live chats, social media threads, or multi-party conversations, this assumption may break, requiring additional preprocessing steps. Social understanding can also be culturally specific, and while one of our datasets is primarily Chinese in origin, more work is needed to incorporate datasets from diverse cultural settings.

Ethical Considerations

Our work raises several important ethical considerations that must be acknowledged to ensure the responsible development and use of socially grounded language models. The model's relational reasoning relies on datasets (e.g., MentalManip, SocialDial) that reflect specific cultural and linguistic contexts, which may introduce cultural or demographic biases (e.g., assumptions about social norms, politeness, or relationships) that do not generalize globally.

Additionally, our work uses data from the SocialDial dataset translated from Chinese to English, which may contain hallucinations. We urge caution in relying on model predictions for sensitive domains such as interpersonal disputes, human resources, or counseling. As probabilistic approximators, language models cannot substitute for nuanced human judgment, particularly in socially or emotionally complex scenarios.

References

AI@Meta. 2024. The llama 3 herd of models. https://ai.meta.com/research/publications/the-llama-3-herd-of-models/.

Sven Buechel, Anneke Buffone, Barry Slaff, Lyle Ungar, and João Sedoc. 2018. Modeling empathy and distress in reaction to news stories. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 4758–4765, Brussels, Belgium. Association for Computational Linguistics.

Yohan Jo Chen, Rui Zhao, and Smaranda Muresan. 2020. Detecting strategy tactics in persuasion dialogues. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 4689–4695.

Marta R. Costa-jussà, James Cross, Onur Celebi Wang, et al. 2022. No language left behind: Scaling human-centered machine translation. arXiv preprint arXiv:2207.04672.

Erenay Dayanik. 2022. Challenges of computational social science analysis with NLP methods. Ph.D. thesis, Dissertation, Stuttgart, Universität Stuttgart, 2022.

DeepMind@Google. 2024. Gemma 2: Improving open language models at a practical size.

Ritam Dutt, Zhen Wu, Jiaxin Shi, Divyanshu Sheth, Prakhar Gupta, and Carolyn Rose. 2024. Leveraging machine-generated rationales to facilitate social meaning detection in conversations. In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 6901–6929, Bangkok, Thailand. Association for Computational Linguistics.

Alan P Fiske and Nick Haslam. 1996. Social cognition is thinking about relationships. Current directions in psychological science, 5(5):143–148.

Chris D Frith. 2007. The social brain? Philosophical Transactions of the Royal Society B: Biological Sciences, 362(1480):671–678.

Haoyu Gao, Ting-En Lin, Hangyu Li, Min Yang, Yuchuan Wu, Wentao Ma, and Yongbin Li. 2023. Self-explanation prompting improves dialogue understanding in large language models. arXiv preprint arXiv:2309.12940.

Christopher Hidey, Elena Musi, Alyssa Hwang, Smaranda Muresan, and Kathy McKeown. 2017. Analyzing the semantic types of claims and premises in an online persuasive forum. In Proceedings of the 4th Workshop on Argument Mining, pages 11–21, Copenhagen, Denmark. Association for Computational Linguistics.

Robert A Hinde. 1976. Interactions, relationships and social structure. Man, pages 1–17.

Robert A Hinde. 1987. Individuals, relationships and culture: Links between ethology and the social sciences. CUP Archive.

Yuxin Hou and Junming Huang. 2025. Natural language processing for social science research: A comprehensive review. Chinese Journal of Sociology, 11(1):121–157.

Itay Itzhak, Gabriel Stanovsky, Nir Rosenfeld, and Yonatan Belinkov. 2024. Instructed to bias: Instruction-tuned language models exhibit emergent cognitive bias. Transactions of the Association for Computational Linguistics, 12:771–785.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b.

Shafiq Joty, Giuseppe Carenini, Raymond Ng, and Gabriel Murray. 2019. Discourse analysis and its applications. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts, pages 12–17.

Tushar Khot, Harsh Trivedi, Matthew Finlayson, Yao Fu, Kyle Richardson, Peter Clark, and Ashish Sabharwal. 2023. Decomposed prompting: A modular approach for solving complex tasks. In The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023. OpenReview.net.

Jaekyeom Kim, Dong-Ki Kim, Lajanugen Logeswaran, Sungryull Sohn, and Honglak Lee. 2024. Auto-intent: Automated intent discovery and self-exploration for large language model web agents. In Findings of the Association for Computational Linguistics: EMNLP 2024, pages 16531–16541, Miami, Florida, USA. Association for Computational Linguistics.

Bolei Ma, Yuting Li, Wei Zhou, Ziwei Gong, Yang Janet Liu, Katja Jasinskaja, Annemarie Friedrich, Julia Hirschberg, Frauke Kreuter, and Barbara Plank. 2025a. Pragmatics in the era of large language models: A survey on datasets, evaluation, opportunities and challenges. arXiv preprint arXiv:2502.12378.

Jiayuan Ma, Hongbin Na, Zimu Wang, Yining Hua, Yue Liu, Wei Wang, and Ling Chen. 2025b. Detecting conversational mental manipulation with intent-aware prompting. In Proceedings of the 31st International Conference on Computational Linguistics, pages 9176–9183, Abu Dhabi, UAE. Association for Computational Linguistics.

Navonil Majumder, Soujanya Poria, Devamanyu Hazarika, Rada Mihalcea, Alexander Gelbukh, and Erik Cambria. 2019. Dialoguernn: An attentive rnn for emotion detection in conversations. In Proceedings of the AAAI conference on artificial intelligence, volume 33, pages 6818–6825.

Todd Morrill, Zhaoyuan Deng, Yanda Chen, Amith Ananthram, Colin Wayne Leach, and Kathleen McKeown. 2024. Social orientation: A new feature for dialogue analysis. In International Conference on Language Resources and Evaluation.

Soujanya Poria, Erik Cambria, Devamanyu Hazarika, and Devamanyu Vij. 2019. Emotion recognition in conversation: Research challenges, datasets, and recent advances. IEEE Access, 7:100943–100953.

Ofir Press, Muru Zhang, Sewon Min, Ludwig Schmidt, Noah A Smith, and Mike Lewis. 2022. Measuring and narrowing the compositionality gap in language models. arXiv preprint arXiv:2210.03350.

Cheng Qian, Bingxiang He, Zhong Zhuang, Jia Deng, Yujia Qin, Xin Cong, Zhong Zhang, Jie Zhou, Yankai Lin, Zhiyuan Liu, and Maosong Sun. 2024. Tell me more! towards implicit user intention understanding of language model driven agents. In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 1088–1113, Bangkok, Thailand. Association for Computational Linguistics.

Qwen Team. 2024. Qwen2.5: A party of foundation models.

Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 2019. Towards empathetic open-domain conversation models: a new benchmark and dataset. arXiv preprint arXiv:1811.00207.

Disha Sheshanarayana, Tanishka Magar, Ayushi Mittal, and Neelam Chaplot. 2025. Unmasking the strategists: An intent-driven multi-agent framework for analyzing manipulation in courtroom dialogues. In Proceedings of the Third Workshop on Social Influence in Conversations (SICon 2025), pages 97–108, Vienna, Austria. Association for Computational Linguistics.

Weiyan Shi, Liang Qiu, Dehong Xu, Pengwei Sui, Pan Lu, and Zhou Yu. 2023. KokoMind: Can Large Language Models Understand Social Interactions?

Xinran Wang, Zhou Yu, et al. 2023. Detecting mental manipulation in movie dialogues. arXiv preprint arXiv:2305.19020.

Yuxin Wang, Ivory Yang, Saeed Hassanpour, and Soroush Vosoughi. 2024. Mentalmanip: A dataset for fine-grained analysis of mental manipulation in conversations.

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.

Akhila Yerukola, Saujas Vaduguru, Daniel Fried, and Maarten Sap. 2024. Is the pope catholic? yes, the pope is catholic. generative evaluation of non-literal intent resolution in LLMs. In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pages 265–275, Bangkok, Thailand. Association for Computational Linguistics.

Haolan Zhan, Zhuang Li, Yufei Wang, Linhao Luo, Tao Feng, Xiaoxi Kang, Yuncheng Hua, Lizhen Qu, Lay-Ki Soon, Suraj Sharma, Ingrid Zukerman, Zhaleh Semnani-Azad, and Gholamreza Haffari. 2023. Socialdial: A benchmark for socially-aware dialogue systems. In SIGIR '23: Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval, page 2712–2722, New York, NY, USA. Association for Computing Machinery.

Runcong Zhao, Qinglin Zhu, Hainiu Xu, Jiazheng Li, Yuxiang Zhou, Yulan He, and Lin Gui. 2024. Large language models fall short: Understanding complex relationships in detective narratives. In Findings of the Association for Computational Linguistics: ACL 2024.

Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc V. Le, and Ed H. Chi. 2023. Least-to-most prompting enables complex reasoning in large language models. In

The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023. OpenReview.net.

Hao Zhou, Tom Young, Minlie Huang, Haizhou Zhao, Jingfang Xu, and Xiaoyan Zhu. 2018. Commonsense knowledge aware conversation generation with graph attention. In Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18, pages 4623–4629. International Joint Conferences on Artificial Intelligence Organization.

A Prompts Used in Experiments

A.1 GRASP.

Prompts for this method begin by asking the model to characterize the relationship between the speakers before making a prediction.

The prompt for the task of mental manipulation is:

> *I will provide you with a dialogue. Please address the following two tasks carefully and strictly adhere to the instructions:*
> *Task 1: Describe the relationship between the speakers.*
> *Task 2: Based on Task 1, please determine if the dialogue contains elements of mental manipulation, by answering '[a]' for Yes or '[b]' for No.*
> *Your output must end with a single sentence formatted as: "Therefore, the answer is '[a]'" or "Therefore, the answer is '[b]'"*

The prompt for the task of social norm classification is:

> *I will provide you with a norm rule and a dialogue. Please address the following two tasks carefully and strictly adhere to the instructions:*
> *Task 1: Describe the relationship between the speakers.*
> *Task 2: Based on Task 1, the norm rule, and the dialogue, please determine whether the speaker(s) are adhering to or violating the norm. Please answer '[a]' for Violation or '[b]' for Adherence.*
> *Your output must end with a single sentence formatted as:*
> *"Therefore, the answer is '[a]' " or "Therefore, the answer is '[b]' "*
> {{Norm_Type}}
> {{Norm_Context}}
> {{Norm_Action}}
> {{Dialogue}}

A.2 GRASP_Ext.

This extension of GRASP asks for both relationship and typical level of persuasion in that relationship before asking for the final classification label. The prompt for the task of mental manipulation is:

> *I will provide you with a dialogue. Please address the following three tasks carefully and strictly adhere to the instructions:*
> *Task 1: Describe the relationship between the speakers.*
> *Task 2: Describe whether the dialogue contains a typical level of persuasion or influence that is common in everyday conversations.*
> *Task 3: Based on Task 1 and 2, please determine if the dialogue contains elements of mental manipulation, by answering '[a]' for Yes or '[b]' for No.*
> *Your output must end with a single sentence formatted as: "Therefore, the answer is '[a]'" or "Therefore, the answer is '[b]'"*

A.3 Direct.

This prompt directly asks the model to perform the classification task without asking for the speaker relationship.
The prompt for the task of mental manipulation is:

> *I will provide you with a dialogue. Please determine if it contains elements of mental manipulation. Just answer with '[a]' for Yes or '[b]' for No, and don 't add anything else. Your output must end with a single sentence formatted as:*
> *"Therefore, the answer is '[a]'" or "Therefore, the answer is '[b]'"*

The prompt for the task of social norm classification is:

> *I will provide you with a norm rule and a dialogue. Based on the norm rule and the dialogue, please determine whether the speaker(s) are adhering to or violating the norm. Please answer '[a]' for Violation or '[b]' for Adherence and don 't add anything else. Your output must end with a single sentence formatted as: "Therefore, the answer is '[a]' " or "Therefore, the answer is '[b]' "*
>
> {{Norm_Type}}
> {{Norm_Context}}
> {{Norm_Action}}
> {{Dialogue}}

A.4 Direct_Wang.

This is the prompt used by Wang et al. (2024) for the task of mental manipulation. The prompt is as follows:

> *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.*

A.5 IAP

This is a method proposed by Ma et al. (2025b) for the task of mental manipulation. The prompt is as follows:

> *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.*
> *{{Dialogue}}*
> *{{Person1's intent }}*
> *{{Person2's intent}}*

A.6 CoT

This is the Chain-of-Thought prompting baseline proposed by Wei et al. (2022) to elicit step-by-step reasoning before producing a final answer. The prompt is as follows:

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

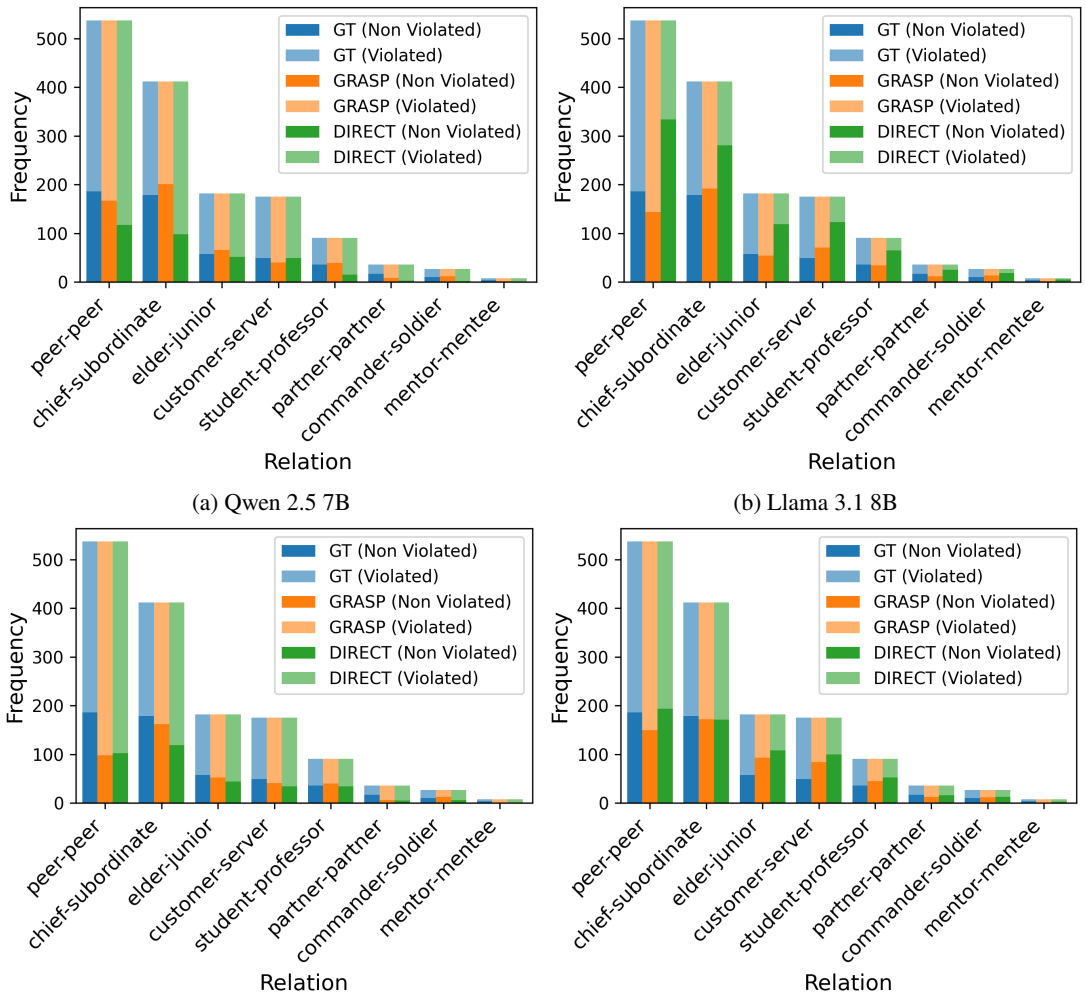
A.7 Self-Explain

This is the self-explanation prompting baseline proposed by Gao et al. (2023), where the model explains each utterance in the dialogue before making a final prediction. The prompt is as follows:

> *I will provide you with a dialogue. For each utterance, briefly explain what the speaker means or intends. After completing the explanations, determine if the dialogue contains elements of mental manipulation. Finally, just answer with 'Yes' or 'No', and don't add anything else.*

A.8 Self-Ask

This is the self-ask prompting baseline proposed by Press et al. (2022), where the model generates and answers sub-questions before making the final prediction. The prompt is as follows:



(a) Qwen 2.5 7B

(b) Llama 3.1 8B

(c) Gemma-2 9B

(d) Mistral 7B

Figure 4: Label distribution across relationship types in the SocialDial dataset (social norm classification task). Each cluster of bars corresponds to a gold relationship (as annotated in the dataset), with bar colors indicating the method and class label. Color shading distinguishes between violated or adhered labels within each method. GRASP's predicted distribution better aligns with the ground truth distribution compared to DIRECT

.

> *I will provide you with a dialogue. First, generate any follow-up questions you need in order to decide whether the dialogue contains mental manipulation. Answer each of these sub-questions. Then, based on your answers, determine if the dialogue contains elements of mental manipulation. Finally, just answer with 'Yes' or 'No', and don't add anything else.*

B Impact of Relationship Type

B.1 Normalized Relation-wise Gains

To better quantify the source of the gains while controlling for data imbalance, we define the following normalized gain metric:

$$\Delta F1(r) = \frac{(F1_{\text{GRASP},r} - F1_{\text{DIRECT},r}) \times \text{freq}_r}{\sum_r \text{freq}_r}$$

where freq_r corresponds to the frequency of the relationship r and $F1_{m,r}$ corresponds to the F1 score of method m on instances belonging to relation r. Figures 8 and 9 show the normalized F1 scores per relation in the MentalManip and SocialDial datasets.

C Comparison in tokens and latency

We also perform a comparison of input tokens (context size due to prompts), output tokens, and inference time across multi-step inference methods (GRASP, IAP, CoT, and Self-Explanation) and the single-step baseline (DIRECT). On average, multi-step methods add only 15 more input tokens and produce 110–115 additional output tokens compared to single-step. In terms of inference time, single-step methods require 1.5–2 minutes per sample, while multi-step methods range between 5–7 minutes. Importantly, the token and time differences among multi-step methods themselves are insignificant, and since GRASP consistently outperforms all other approaches, it offers the best performance–time trade-off overall.

D Implementation Details

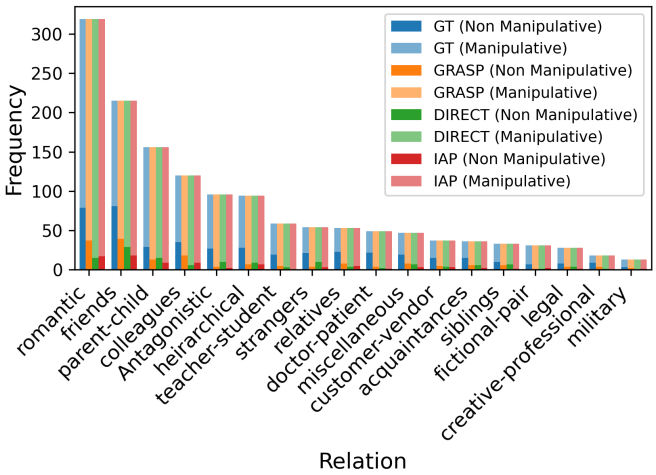
We employ LLMs built using the HuggingFace Transformers library[2], with vLLM[3] enabling efficient inference and PyTorch[4] serving as the underlying deep learning framework. All experiments are conducted on a single NVIDIA A100 GPU. For all the models, we set temperature to 0.6 to balance diversity and determinism in outputs. To account for variability in generation, we repeat each experiment three times and report average performance across runs. Our experiments utilize LLaMA-3.1-8B[5], Qwen-2.5-7B[6], Mistral-7B[7], Gemma-9B[8], and a quantized version of LLaMA-3.1-70B[9]. We additionally employ NLTK[10] for tokenizing dialogues in experiments requiring dialogue-level analysis (e.g., Figure 3).

For the MentalManip dataset, we use data sourced from a public repository[11], which is distributed under the CC BY-NC 4.0 license. This is in line with the intended use.

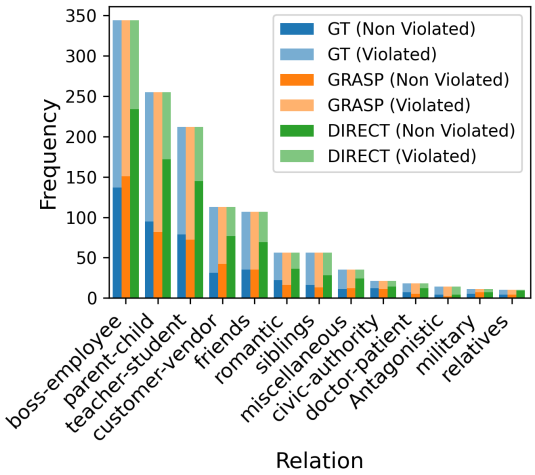
Each run for 7B–9B parameter models required approximately 1 GPU hour on average, while experiments with LLaMA-70B took 3–4 GPU hours per run.

In this work, we used AI assistants for minor grammatical corrections while writing the draft.

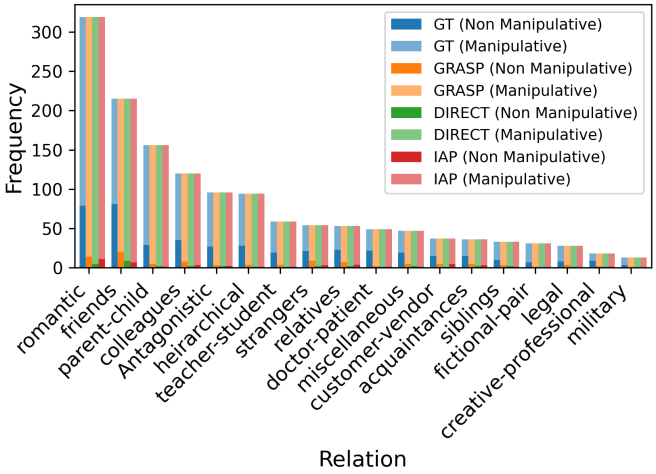
[2] https://huggingface.co/
[3] https://github.com/vllm-project/vllm
[4] https://pytorch.org/
[5] https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct
[6] https://huggingface.co/Qwen/Qwen2.5-7B-Instruct
[7] https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.3
[8] https://huggingface.co/google/gemma-2-9b-it
[9] https://huggingface.co/RedHatAI/Meta-Llama-3.1-70B-Instruct-quantized.w8a8
[10] https://www.nltk.org/
[11] https://github.com/audreycs/MentalManip



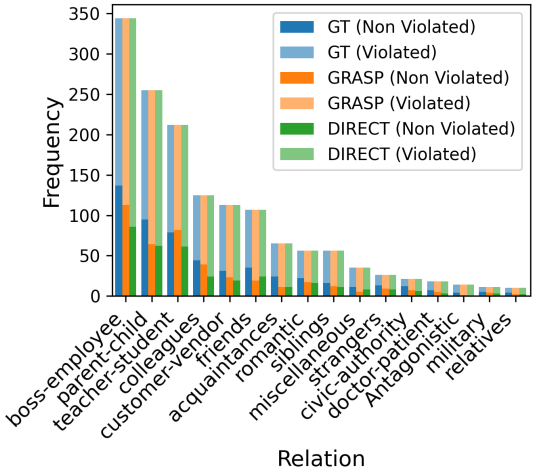
(a) Llama 3.1 8B



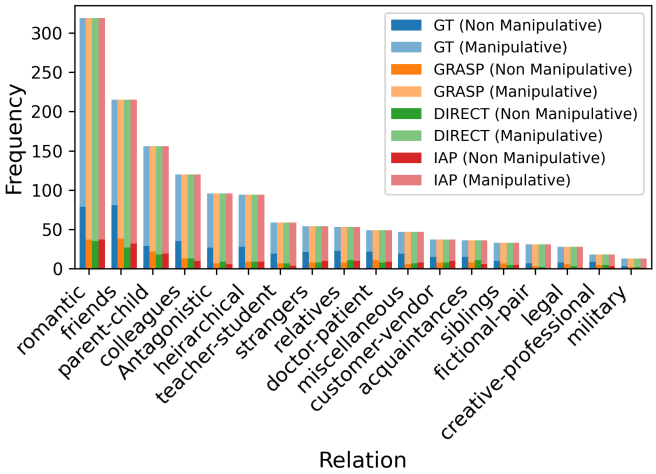
(b) Llama 3.1 8B



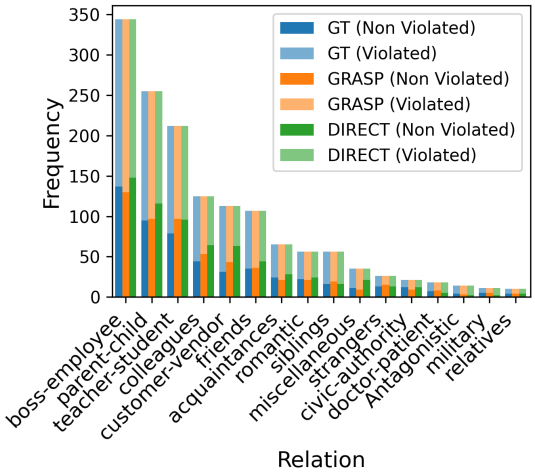
(c) Gemma 9B



(d) Gemma 9B

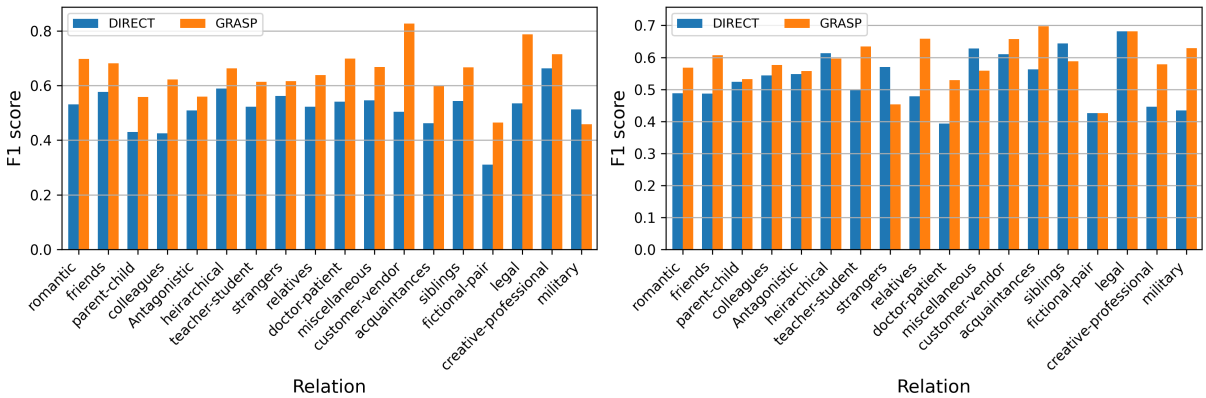


(e) Mistral 7B
MentalManip

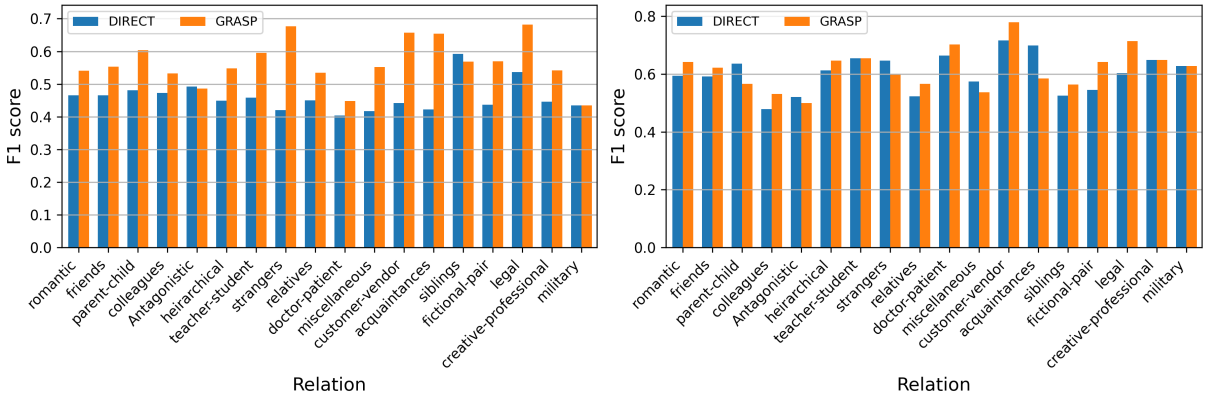


(f) Mistral 7B
SocialDial

Figure 5: Label distribution for different relations for different LLMs. The charts on the left are for the mental manipulation task and those on the right are for social norm classification task. Each cluster of bars corresponds to a specific relation, with bar colors indicating the method and class label. Color shading distinguishes between manipulative (or violated) and non-manipulative (or adhered) outcomes within each method. GRASP's predicted distributions better align with the ground truth distributions compared to all baselines.



(a) Qwen 2.5 7B

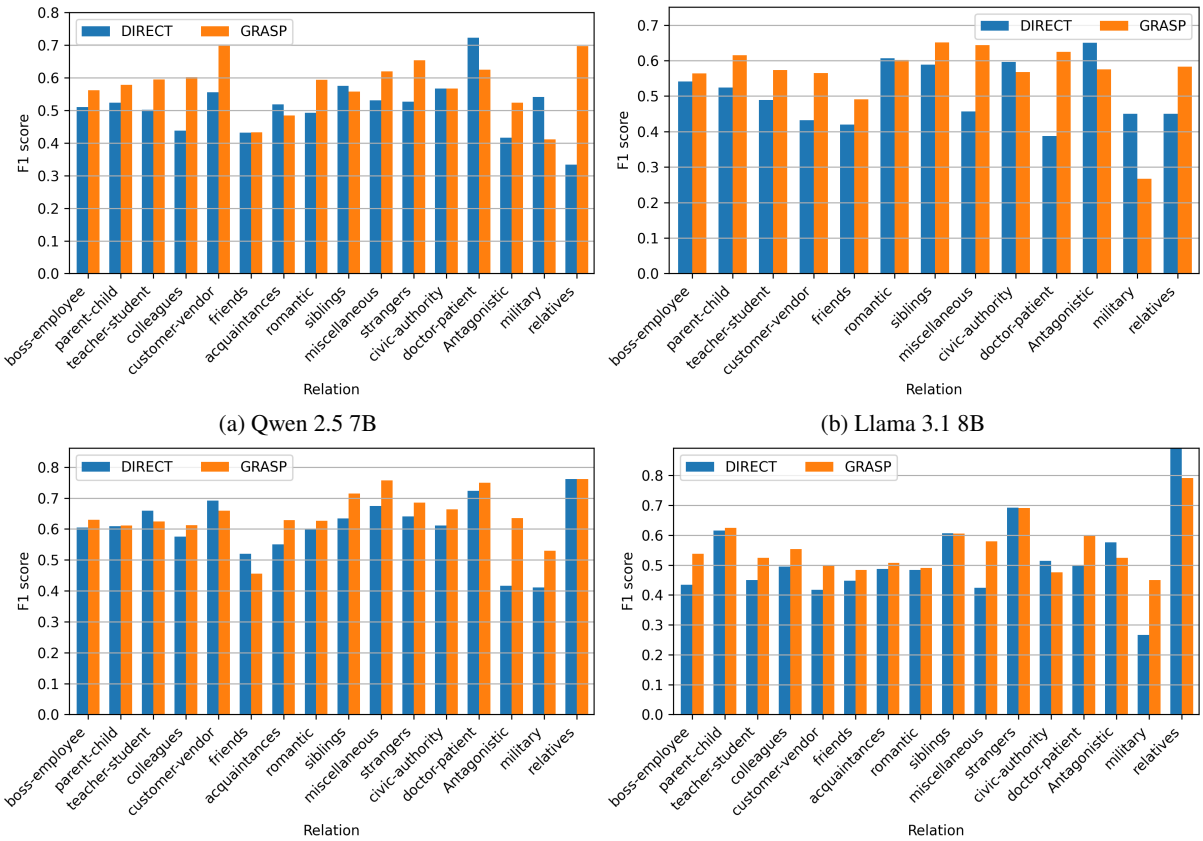


(b) Llama 3.1 8B

(c) Gemma-2 9B

(d) Mistral 7B

Figure 6: Mental Manipulation Task. Macro F1 scores per relationship for DIRECT and GRASP. The y-axis indicates F1 score, and the x-axis shows relations. We can observe that GRASP (orange) yields significant gains over the DIRECT (blue) for almost all relations.



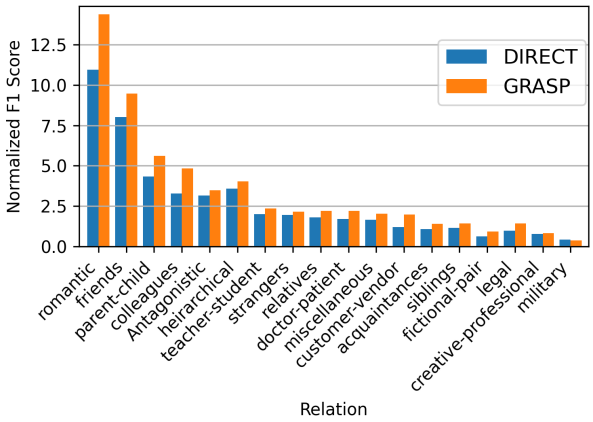
(a) Qwen 2.5 7B

(b) Llama 3.1 8B

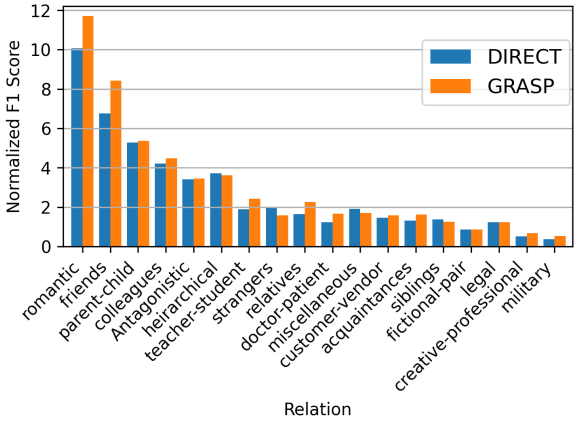
(c) Gemma-2 9B

(d) Mistral 7B

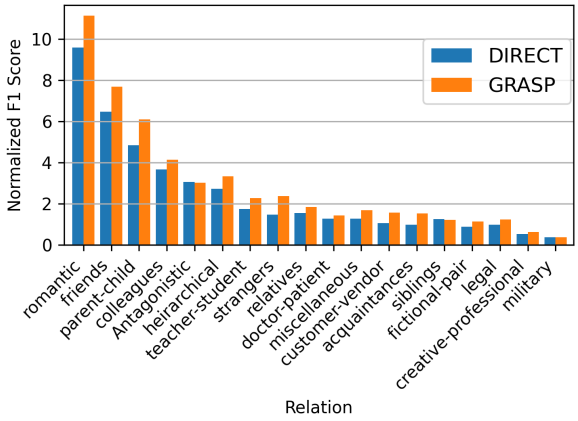
Figure 7: Social Norm Classification Task. Macro F1 scores per relationship for DIRECT and GRASP. The y-axis indicates F1 score, and the x-axis shows relations. We can observe that GRASP (orange) yields significant gains over the DIRECT (blue) for almost all relations.



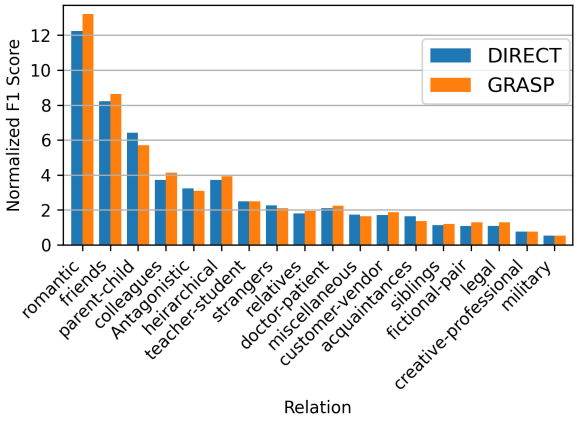
(a) Qwen 2.5 7B



(b) Llama 3.1 8B

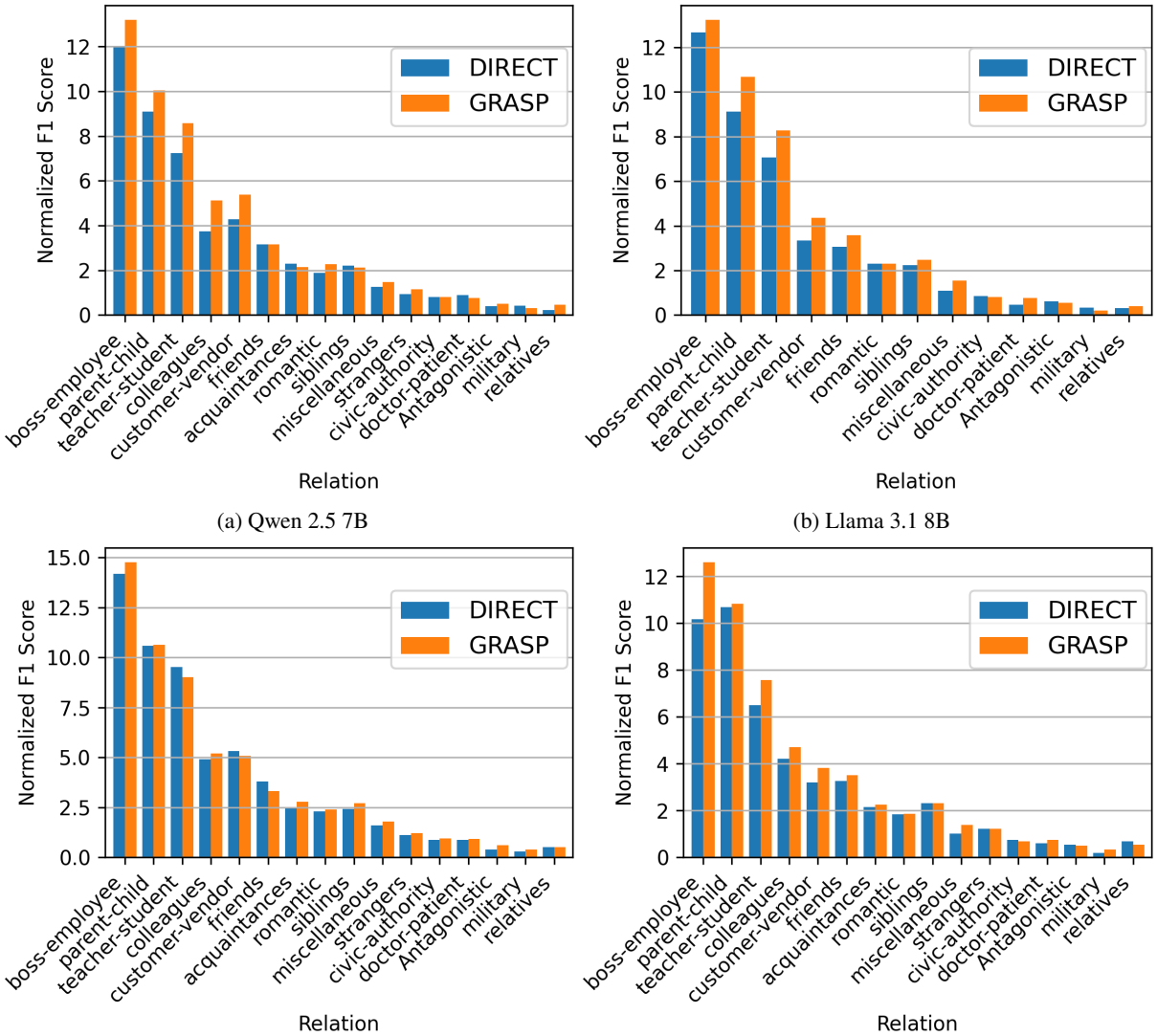


(c) Gemma-2 9B



(d) Mistral 7B

Figure 8: Mental Manipulation Detection Task. Normalized Macro F1 scores per relationship for DIRECT and GRASP. The y-axis indicates F1 score, and the x-axis shows relations. We can observe that GRASP (orange) yields significant gains over the DIRECT (blue), mainly in symmetric relationships.



(a) Qwen 2.5 7B

(b) Llama 3.1 8B

(c) Gemma-2 9B

(d) Mistral 7B

Figure 9: Social Norm Classification Task. Normalized Macro F1 scores per relationship for DIRECT and GRASP. The y-axis indicates F1 score, and the x-axis shows the relations. We can observe that GRASP (orange) yields significant gains over the DIRECT (blue), mainly in asymmetric relationships.

Mental Manipulation Detection Task	
Example 1	
Dialogue	Person1: Nobody else is going down but me, okay? Person2: Hey. Bein' ballsy with your life now doesn't change what came before – it's just stupid. Person1: What, you think I'm doin' this to prove something? Person2: You said it, not me. Let's just not let one bad call lead to another. Person1: Thanks for the tip, Johns.
Gold Label	Non Manipulative
LLAMA 3.1 8B - DIRECT	Manipulative. Person 2 is trying to influence Person 1's behavior by using a gentle yet cautionary tone. Person 2 is attempting to prevent Person 1 from making another "bad call" by implying that Person 1's current actions are impulsive. This is an example of mental manipulation, as Person 2 is attempting to steer Person 1 away from a certain course of action without directly ordering them to do so.
+ GRASP	Non Manipulative. *Inferred Relationship:* Two friends or colleagues. The relationship between the speakers appears to be between two friends or colleagues, as they use informal language and familiar nicknames. There seems to be a history between them, as Person 2 references "one bad call" that Person 1 is trying to rectify. Based on the relationship established, the dialogue does not contain overt elements of mental manipulation, as Person 2 is speaking from a place of concern and trying to offer advice rather than manipulating Person 1's thoughts or actions.
Example 2	
Dialogue	Person1: This better not be a wild goose chase. Person2: Or what, you'll authorize my death a second time today? Person1: Don't forget, convict, if this psycho isn't stopped, you go right back to rotting in a prison cell.
Gold Label	Manipulative
QWEN 2.5 7B - DIRECT	Non Manipulative. The dialogue contains an implicit threat and a form of emotional manipulation through the use of a convict's fear of incarceration. However, the primary context is a conversation about stopping a "psycho," and the manipulation is more situational rather than direct mental manipulation. Therefore, it is non-manipulative.
+ GRASP	Manipulative. *Inferred Relationship:* Law enforcement-Prisoner The relationship between the speakers appears to be that of a law enforcement officer (Person1) and a prisoner (Person2). There is a tone of authority and tension in the dialogue, suggesting a power dynamic where Person 1 is in charge and Person 2 is under their control. Given the power dynamic and the threatening tone, there are elements of mental manipulation present, as Person 1 is using authority and the threat of returning Person 2 to a prison cell to influence their behavior or decisions.
Example 3	
Dialogue	Person1: Well don't you? Person2: I think you should try not to think about this anymore. Person1: Really? Person2: Yeah. Person1: Oh. Okay. I'll try that then.
Gold Label	Non Manipulative
GEMMA-2 9B - DIRECT	Manipulative. Person 2 is attempting to control Person 1's thoughts and feelings by suggesting they stop thinking about something. This is a form of manipulation.
+ GRASP	Non Manipulative. *Inferred Relationship:* Friends, Family, or Acquaintances The relationship between the speakers is unclear from the dialogue. They could be friends, family, or acquaintances. Person 2 seems to be offering advice or trying to comfort Person 1. Therefore, it is non-manipulative.

Example 4	
Dialogue	Person1: Nothing. I'm telling you, I just want to make you happy. That's all. Just trust me. Person2: I don't want you to hurt me. Person1: I don't want to hurt you. I'll be gentle. Person2: Do you care about me? Person1: Of course I do.
Gold Label	Manipulative
Mistral 7B - Direct	Non Manipulative. The dialogue does not necessarily contain elements of mental manipulation. The conversation seems more focused on reassuring Person 2 about their concerns, rather than manipulating them.
+ GRASP	Manipulative. *Inferred Relationship:* Romantic or Intimate relationship. The speakers in this dialogue are in a romantic or intimate relationship, as Person 1 is trying to reassure Person 2 about their intentions and care for them. Based on the relationship, there are elements of mental manipulation in this conversation. Person 1 is using various tactics to influence Person 2, such as repeating the phrase "I just want to make you happy" and promising to be gentle, which can be seen as an attempt to make Person 2 feel more comfortable and trusting.

Social norm classification task

Example 1	
Dialogue	Person1: This ring was bought last month, can it be repaired? Person2: Yes. Person1: Thank you, so should I repair it? Person2: We usually have to wait about three days. Person1: Can I mail it? Person2: No problem. Norm Type: Greetings Norm Action: It is obligatory to acknowledge the customer with a standard greeting (words or smile/bow), but it is not necessary to use words or have extensive interaction beyond what is required for the transaction; it is acceptable to get right to the point of the interaction. Norm Context: A public setting, one person is serving the other in a professional capacity.
Gold Label	Violated
Llama 3.1 8B - Direct	Adhered. Person 2 is not using a standard greeting to acknowledge the customer, but is still providing a necessary response to the customer's inquiry. However, the lack of a greeting is not a significant breach of the norm in this context.
+ GRASP	Violated. *Inferred Relationship:* Customer-Service Provider The relationship between the speakers is that of a customer (Person 1) and a service provider (Person 2) in a professional capacity, likely at a store or a repair service. Person 2 is not acknowledging Person 1 with a standard greeting, which is obligatory in a public setting where they are serving the customer in a professional setting.

Example 2	
Dialogue	Person1: I came first, of course, this is my place. What do you mean, you're taking this book? Person2: I'm sorry, I didn't see it. I'm leaving right now. Norm Type: Apology Norm Action: The person should give a light apology Norm Context: A minor offence has been committed
Gold Label	Adhered
QWEN 2.5 7B - DIRECT	Violated. Person 2 says "I'm sorry, I didn't see it, I'm leaving right now," which is a light apology for the potential misunderstanding. However, Person 1's response is confrontational and does not reflect an apology or attempt to resolve the situation amicably. The dialogue does not fully meet the norm of a light apology for a minor offense.
+ GRASP	Adhered. *Inferred Relationship:* Casual Acquaintances or Possibly Strangers The relationship between the speakers appears to be casual acquaintances or possibly strangers. Person 2 has acknowledged the minor offense by apologizing, which aligns with the norm of giving a light apology for a minor offense. Person 1 does not seem to require an apology based on the context provided, as the apology is directed towards Person 2.
Example 3	
Dialogue	Person1: Xiahou, come here. Person2: Chairman, are you calling me? What do you need me to do? Person1: Xiahou, please prepare our books, the IRS is ready to come check. Norm Type: Request Norm Action: Use a direct command to make requests Norm Context: A higher status/power party and a lower status/party who are already familiar with each other
Gold Label	Adhered
GEMMA-2 9B - DIRECT	Violated. Person 2 is adhering to the norm by responding politely and asking what is needed. Person 1 is violating the norm by using a direct command instead of a more indirect request
+ GRASP	Adhered. *Inferred Relationship:* Higher status-Lower status The speakers are a higher status/power party (Person1, referred to as "Chairman") and a lower status/power party (Person2, referred to as "Xiahou"). Person 1 used a direct command ("come here"), which aligns with the norm of using direct commands in requests between a higher and lower status individual.

Example 4	
Dialogue	Person1: Professor, I want to apply for you to be my mentor. Person2: You can fill in the system first. Person1: So you're promising or not? Person2: This is how you ask teachers, is this how graduate students treat teachers? Norm Type: persuasion Norm Action: via logic, comparisons of advantages and disadvantages, Historical precedence, a chain of syllogisms Norm Context: the process of trying to convince someone to agree to do something
Gold Label	Violated
Mistral 7B - Direct	Adhered. Person 2 is not violating the norm of persuasion. Instead, Person2 is responding in a way that is more indirect and respectful, suggesting a more formal approach to asking for a favor, which is appropriate in an academic setting.
+ GRASP	Violated. *Inferred Relationship:* Student-Teacher The relationship between speakers appears to be that of a student seeking guidance from a teacher. Person 1 is asking for a favor, while Person 2 is acting as the authority figure who has the power to grant or deny the request. Based on this, Person 1 is violating the norm. The student is attempting to persuade Person 2 using a more informal and pushy approach, which is not in line with the norm action of using logic, comparisons of advantages and disadvantages, historical precedence, or a chain of syllogisms to persuade.

Table 6: Examples, reasoning, and classification with and without GRASP for Mental manipulation detection and social norm violation tasks.

	Category 1: Relationship inferred incorrectly
Task	Social norm classification
Dialogue	Person1: Hey, can you send me the report right now? Person2: I was planning to share it this evening. Person1: No, I need it immediately. Don't delay. Person2: You don't have to order me around. Person1: I'm just making sure deadlines are respected. Person2: Still, you could ask more politely. Norm Type: Politeness Norm Action: Make requests respectfully among peers Norm Context: Workplace peer-to-peer interaction
Gold Label	Violated
MISTRAL 7B - DIRECT + GRASP	Adhered. *Inferred Relationship:* Manager–Subordinate. The model misinterprets the relationship as hierarchical and therefore judges Person 1's direct commands as acceptable. This incorrect inference leads to an erroneous prediction of adherence to the politeness norm.
	Category 2: Relationship excuses behavior
Task	Mental manipulation detection
Dialogue	Person1: If you cancel on me again, I'll be so hurt. Person2: I told you, I already promised my parents I'd help tonight. Person1: Don't you care about me at all? Person2: That's not fair, of course I do. Person1: Then prove it and stay with me. Person2: . . . I don't know what to say.
Gold Label	Manipulative
LLAMA 3.1 70B - DIRECT + GRASP	Non Manipulative. *Inferred Relationship:* Romantic partners. The model interprets Person 1's emotionally charged requests as typical expressions within an intimate relationship. As a result, it excuses the behavior and fails to identify the manipulative intent present in the dialogue.
	Category 3: Relationship is uncertain
Task	Mental manipulation detection
Dialogue	Person1: Just do it. Person2: Why should I? Person1: Because I told you to. Person2: That's not a reason.
Gold Label	Non Manipulative
LLAMA 3.1 8B - DIRECT + GRASP	Manipulative. *Inferred Relationship:* Superior–Subordinate (uncertain). The model assumes a hierarchical dynamic where Person 1 holds authority over Person 2. Based on this inferred relationship, Person 1's commands are judged as manipulative, even though the short and ambiguous dialogue provides insufficient context to support this inference.

Table 7: Error analysis for GRASP showing the most common categories of errors across models and tasks.