

Are LLMs Effective Negotiators? Systematic Evaluation of the Multifaceted Capabilities of LLMs in Negotiation Dialogues

Anonymous ACL submission

Abstract

A successful negotiation demands a deep comprehension of the conversation context, Theory-of-Mind (ToM) skills to infer the partner’s motives, as well as strategic reasoning and effective communication, making it challenging for automated systems. Given the remarkable performance of LLMs across a variety of NLP tasks, in this work, we aim to understand how LLMs can advance different aspects of negotiation research, ranging from designing dialogue systems to providing pedagogical feedback and scaling up data collection practices. To this end, we devise a methodology to analyze the multifaceted capabilities of LLMs across diverse dialogue scenarios covering all the time stages of a typical negotiation interaction. Our analysis adds to the increasing evidence for the superiority of GPT-4 across various tasks while also providing insights into specific tasks that remain difficult for LLMs. For instance, the models correlate poorly with human players when making subjective assessments about the negotiation dialogues and often struggle to generate responses that are contextually appropriate as well as strategically advantageous.

1 Introduction

As AI systems grow increasingly prevalent, they must be capable of comprehending and participating in strategic conversations that necessitate the consideration of multiple, often conflicting, goals (Chawla et al., 2023b). For instance, AI teammates need to consider both productivity and teamwork; tools for therapists must take both clinical objectives and rapport with the patients into account. Additionally, AI negotiators must effectively balance their own requirements with the needs of their partners to reach a favorable agreement while also ensuring a positive relationship with their partners.

Figure 1 summarizes four core capabilities that are essential for success in a typical bilateral multi-issue negotiation (Fershtman, 1990). The play-

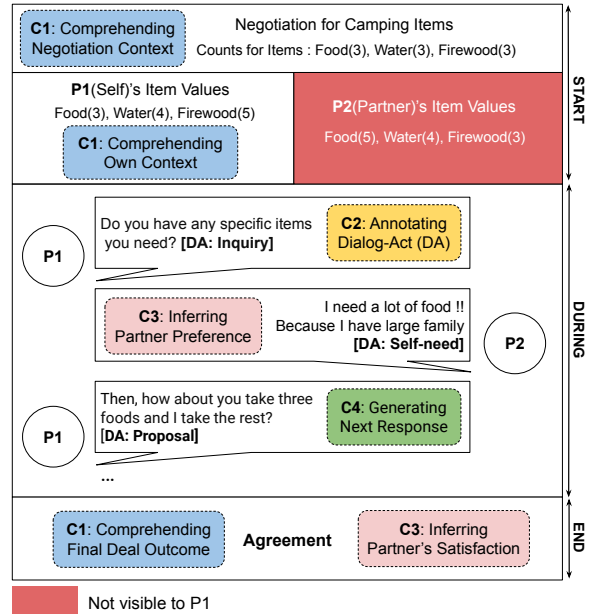


Figure 1: Overview of the key capabilities (C1-C4) required for a successful negotiation. We design tasks aligned with these abilities to assess how LLMs can advance different aspects of negotiation research. The negotiation scenario is based on Chawla et al. (2021b).

ers start with predefined value preferences towards some issues (e.g., food, water, and firewood) and then engage in a dialogue to divide all the available items among each other with the goal of getting the maximum total value out of the negotiation. To achieve success, a negotiation agent must have a deep comprehension of its initial context as well as an understanding of how the negotiation concluded (C1). During the negotiation, it must also understand the structure (e.g., intents and slots) and the semantics (e.g., employed linguistic strategies) of the utterances (C2). Notice that the partner’s context is not visible to the agent at the start of the conversation – it is only through the dialogue that the agent must infer its partner’s preferences so as to come to an agreement with them (C3). Finally,

the agent must be able to generate a coherent and strategically appropriate response conditioned on the input context and conversation history (C4). These complexities make negotiation a challenging testbed for AI research.

Large Language Models (LLMs) have demonstrated remarkable performance in various text comprehension and reasoning tasks (Brown et al., 2020; Kojima et al., 2023). They have also shown exceptional capabilities as dialogue systems, accurately understanding dynamically changing conversation contexts and generating human-like responses (OpenAI, 2022; Chiang et al., 2023). Consequently, it is only natural to explore how LLMs can be useful in negotiations as well – for designing end-to-end (Lewis et al., 2017) and modular negotiation dialogue systems (He et al., 2018), for providing pedagogical feedback through ToM (Johnson et al., 2019), and for scaling up data collection practices by automating the collection of utterance-level annotations (Yamaguchi et al., 2021). Although some recent efforts have applied LLMs to negotiation dialogues (Fu et al., 2023; Gandhi et al., 2023; Abdelnabi et al., 2023), an analysis of the multifaceted capabilities of LLMs in strategic interactions is missing, leaving uncertainties about the numerous ways in which LLMs can advance negotiation research.

To this end, we design a novel methodology to conduct a systematic analysis of state-of-the-art LLMs in negotiations based on 35 tasks developed on 4 dialogue datasets. Our methodology leverages 3 crucial axes to define these tasks in a principled manner: 1) *Objectivity* (Objective vs. Subjective tasks), 2) *Time Stage* of the negotiation (Start, During, and End), and 3) *Task Type* (Comprehension, Annotation, Partner Modeling, and Generation), aligned with the core competencies necessary for successful negotiations (Figure 1). We hope that the insights gained from our analysis enable the incorporation of LLMs to tackle diverse challenges in this area. We now summarize our contributions:

1. We design a novel methodology to evaluate the multifaceted capabilities of LLMs in negotiations. Although our framework is designed with the goal of supporting negotiation research, the resulting methodology captures the nuances common in other dialogue tasks as well (Section 2).
2. We evaluate zero-shot out-of-the-box capabilities of LLMs on 35 tasks based on 4 dialogue

datasets capturing diverse negotiation scenarios. Our overall trends show the superiority of GPT-4, finding that it often outperforms even the Flan-T5 baseline which is specifically fine-tuned for the corresponding task (Section 4). Additionally, all models show poor correlations with human judgments on subjective assessments about the negotiation.

3. Through a human evaluation, we find that GPT-4 performs comparable to an average crowdsourcing worker in response generation. We perform an error analysis, discussing the coherency and strategic failures that still persist in the compared LLMs (Section 4.4).
4. We study the impact of various prompting strategies and provide recommendations for future work (Sections 5 and 7).

2 Methodology

We evaluate the multifaceted capabilities of LLMs in negotiations with a three-step framework (Figure 2): 1) Designing tasks from human-human dialogue datasets, 2) Constructing prompt templates, and 3) Performing evaluations and analysis across various proprietary and open-source LLMs.

2.1 Datasets

Three key requirements guide our dataset selections: 1) Well-structured design with concrete player preferences and outcomes, 2) Diversity in the task scenario, and 3) Availability of metadata for testing holistic LLM abilities. Given this criteria, we selected 4 datasets: *CRA* (DeVault et al., 2015), *DND* (Lewis et al., 2017), *CA* (Chawla et al., 2021c), and *JJ* (Yamaguchi et al., 2021). Table 1 provides an overview of these datasets. While each dataset captures a distinct scenario, they are all based on a tractable *closed-domain* abstraction from the literature, referred to as Multi-Issue Bargaining Task or MIBT (Fershtman, 1990). MIBT involves bilateral negotiations over fixed quantities of some issues (Table 1). Each player has predefined preferences towards each issue (called *values*), which define the player’s priority order.

The task for a player is to engage in a dialogue with their partner, whose priorities are not known beforehand, and divide all the available items so as to maximize their total points scored – computed based on the assigned preferences and the final agreed deal. MIBT is a generic framework that

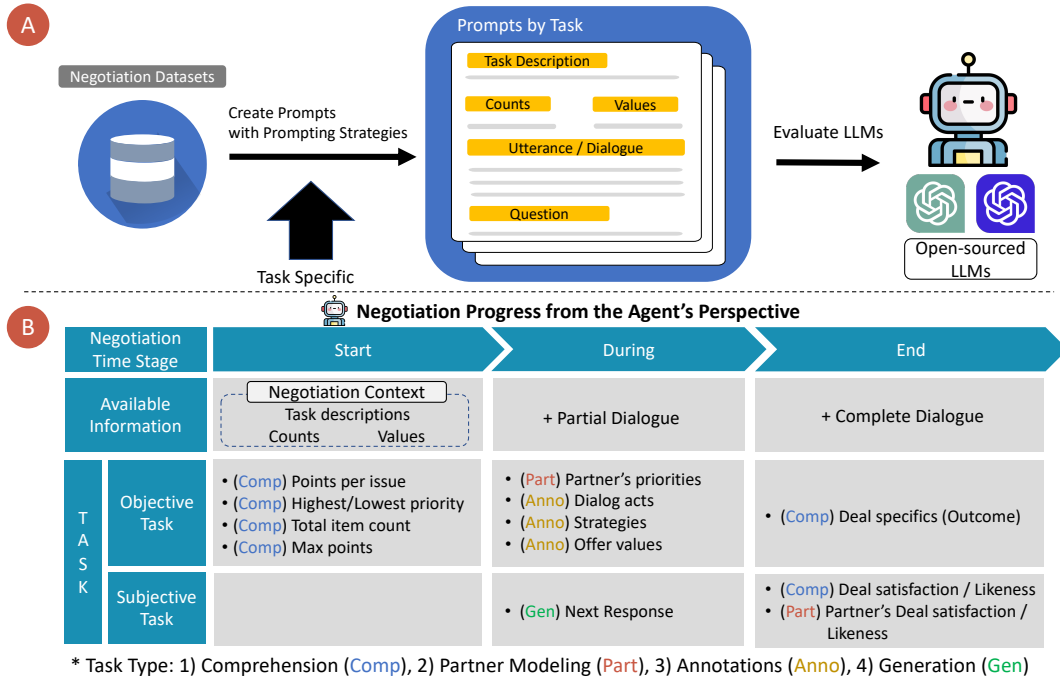


Figure 2: Our methodology for systematically evaluating LLMs in negotiation dialogues. Part A (top) describes the pipeline for creating task-specific prompts from a negotiation dataset and evaluating various LLMs with them. Part B (bottom) depicts the tasks categorized by *Objectivity*, *Time Stage*, and *Task Type* (Section 2.2).

Dataset	Scenario	Issues	Size
CRA	Artifacts trading	(Painting, Lamp, Album)	119
DND	General items	(Ball, Hat, Book)	6,251
CA	Campsite Neighbors	(Food, Water, Firewood)	1,030
JI	Job Recruiter-Worker	(Salary, Day-off, Position, Company, Workplace)	2,639

Table 1: The datasets used in our analysis. We sample 200 test instances for our evaluation. The training data is used for fine-tuning our baseline model.

can capture many interesting real-world negotiations and, hence, has been adopted in NLP, beyond NLP (Mell and Gratch, 2017), and in the industry as well (e.g., iDecisionGames¹).

Some datasets like CA also report subjective metrics like *satisfaction from the outcome* and the player’s *likeness towards their partner*. Some include utterance-level dialogue act (e.g., CRA) and strategy (e.g., CA) annotations. We use this metadata to design a variety of tasks to evaluate LLMs.

2.2 Task Design

We build a total of 35 tasks based on the datasets described above. Our guiding principle is the following: *Every task is designed from the perspective of an agent that is negotiating itself*. This is crucial since it governs the kind of information that we can

use in the prompts for LLMs (e.g., explicit partner preferences are not visible).

To describe our tasks in a principled way, we categorize them based on: 1) *Objectivity*, 2) *Time stage*, and 3) *Task Type* (Figure 2). We discuss these criteria below. Although these are designed with the domain of negotiations in mind, many nuances discussed below directly apply to traditional task-oriented or open-domain dialogues (Wang et al., 2023) as well, making our evaluation framework broadly applicable.

2.2.1 Objectivity

Most tasks are *objective*, meaning that they test whether the models can accurately answer questions on the objective facts about a negotiation. These tasks can be evaluated with standardized metrics from the literature. In contrast, a few tasks are categorized as *Subjective*, where there is no one correct answer. This includes predicting outcome satisfaction and partner likeness as reported by the human players in a post-survey. These subjective measures have been argued to be crucial for successful task outcomes, especially in scenarios involving repeated interactions between the same stakeholders (Oliver et al., 1994; Mell et al., 2019; Chawla et al., 2021a). Since they govern the user’s perception of the agents in human-agent interac-

¹<https://idecisiongames.com/promo-home>

tions, the models must be capable of capturing these effects. Hence, we also design tasks to judge (by computing correlations) whether the model’s predictions align with those reported by the human players. We consider response generation as another subjective task, which is evaluated via automatic and human evaluation.

2.2.2 Time Stage

A negotiation dialogue typically unfolds in 3 stages: *Start*, *During*, and *End*. We make this seemingly trivial distinction since it directly impacts the information used while designing the prompts: in a given stage, we only input the information that an agent would have access to (from its own perspective) at that point in the negotiation.

At the *start* (where no actual dialogue has occurred yet), the agent is only aware of the basic information from the negotiation context, including the scenario description as well as the item counts and priority values for each issue (Figure 2). *During* the dialogue, the LLMs also get access to the partial historical conversation. Finally, tasks defined at the *end* of the negotiation use the entire dialogue within their prompts.

2.2.3 Task Type

We categorize the tasks into four types based on the kind of capabilities they are designed to test.

Comprehension: In the *Start* stage, these tasks assess the model’s ability to understand the negotiation context, which is fundamental for any sensible negotiation agent. This includes asking the total number of available items, the priorities for the issues, and the maximum points the agent can earn from the negotiation. In the *End* stage, Comprehension tasks evaluate whether the models can answer questions about the final outcomes, including details of the final deal and the subjective measures of satisfaction and partner likeness. Proficiency in these tasks can aid in automatically evaluating human-human or human-agent negotiations.

Annotation: These involve detecting dialogue acts (e.g., disagree, propose) or linguistic strategies (e.g., elicit-preference, self-need) in utterances, along with parsing incoming offers. Annotation tasks are critical components of the Natural Language Understanding (NLU) module for a typical modular dialogue system. *During* the negotiation, they are essential in understanding the input from the partner to decide the agent’s future actions.

Partner Modeling: These tasks are closely linked

to Theory of Mind (ToM) (Frith and Frith, 2005), where the goal is to infer the mental states of the partner. In our work, we focus on predicting the partner’s priorities and subjective assessments. Partner Modeling is particularly vital *during* the negotiation, where understanding the other party’s requirements and adapting one’s strategy accordingly can be key for efficient and favorable outcomes.

Generation: Finally, we consider the task of response generation based on context and dialogue history. This includes crafting offers and counteroffers, responding to incoming offers, and other communicative elements that advance the negotiation toward a successful conclusion. This is an essential requirement for a conversational agent and tests whether the models can reason through the current state and respond in a way that is both *contextually appropriate* and *strategically advantageous*.

2.3 Prompting for Evaluation

To assess LLMs’ capabilities on our proposed tasks, we develop prompt templates that are later filled with the relevant information from the dataset instances. The prompts for zero-shot evaluation include (in order) the task description, issue counts & values, the dialogue or an utterance, and the question. For our analysis in Section 5, we also employ alternative prompting strategies, such as few-shot prompting and Chain-of-Thought (CoT) (Wei et al., 2023). We provide task-specific details in Appendix A, including the questions used in the prompts and the dataset applicable for each task. We also provide example prompts in Appendix B.

3 Experiment Design

Our primary goal is to analyze the effectiveness of state-of-the-art LLMs for strategic negotiation interactions. Here, we discuss the compared models and evaluation metrics for the results presented in Section 4 (overall zero-shot results) and Section 5 (task-specific analysis to gain further insights).

3.1 Baselines

We use the Majority-label model and a fine-tuned Flan-T5-base (Chung et al., 2022) as baselines. We selected Flan-T5 for its superior performance across a wide range of NLP tasks and flexibility to handle diverse input and output formats. For fine-tuning, we use the AdamW optimizer (Loshchilov and Hutter, 2019) with an initial learning rate of $5e-5$ and a linear scheduler. We train the models for

a maximum of 5 epochs with a batch size of 8, picking the best-performing checkpoint for evaluation. We provide detailed information about this training process in Appendix C, including data processing steps and the computational resources used.

3.2 LLMs for evaluation

We compare a variety of LLMs, top-performing on popular leaderboards at the time of experimentation. Among proprietary models, we choose OpenAI’s GPT-3.5-Turbo (OpenAI, 2022) and GPT-4 (OpenAI, 2023). Among open-source LLMs, we experiment with Mistral-7B (Jiang et al., 2023) and Wizard-13B (Xu et al., 2023), along with Vicuna 13B and 33B variants (Chiang et al., 2023).

3.3 Evaluation Metrics

The metrics depend on the kind of task being evaluated. For annotation tasks, we use Macro-F1 to account for label imbalance. For subjective tasks that measure satisfaction and likeness scores, we use correlations with human judgments. For response generation, we report both automatic evaluation (BLEU and ROUGE) and human evaluation (response coherency and strategic appropriateness). For all others, primarily comprehension and partner modeling tasks, we rely on Accuracy to measure the performance. The metric used for each task can also be found in Table 6 of Appendix A.

4 Results

Figure 3 and Table 2 present the key overall trends for zero-shot evaluation. We provide the complete task-wise results in Appendix A and discuss the key findings below.

4.1 Comprehension Tasks

As evident in Figure 3a, GPT-4 outperforms all other models on Comprehension tasks in the *Start* stage with an average accuracy of 81%. This was followed by GPT-3.5 and Mistral7B, both still performing above 70%. However, note that these tasks primarily involve questions about understanding the negotiation context given explicitly in the prompt. Hence, a simple rule-based parser would achieve 100% performance for these tasks – indicating a huge room for improvement.

The tasks in the *End* stage are more challenging, requiring the models to comprehend the dialogue, perform reasoning to extract the agreed deal, and then compute the answer. As per Figure 3b, most models fail to perform well, only marginally better

than the trivial Majority baseline. GPT-4 still exhibits superior performance, even surpassing Flan-T5. In comparison, since all evaluated instances end in an *unambiguous* final deal, an expert human can achieve nearly a perfect score on these tasks.

For subjective tasks, we analyze if the LLMs’ self-assessments of outcome satisfaction and likeness for the partner align with those reported by the human players. The results from Figure 3c show that this is not the case, with the models performing poorly on Accuracy and PCC. This leaves uncertainties about the LLMs’ abilities to capture the psychological states of the human players in negotiations. Future work on understanding the psychological personalities exhibited by LLMs can help to explain these observations.

4.2 Annotation Tasks

Figure 3d reveals that both GPT-3.5 and GPT-4 beat the fine-tuned Flan-T5 baseline on annotation tasks, achieving 62.4% and 52.5% Macro-F1 respectively. Task-wise results (Table 6 of Appendix A) show that these trends are driven by the poor performance of Flan-T5 on *dur_dial_act_ji* task, but on most tasks, Flan-T5 performs better than GPT-4, indicating room for improvement in zero-shot scenarios. Further, models find it more challenging to detect negotiation strategies than dialogue acts, probably because the strategies are often expressed in a more subtle manner. While these results are without few-shot examples or prior utterances – we will explore these possibilities later in Section 5.

4.3 Partner Modeling Tasks

Figure 3e shows the results for the objective partner modeling tasks (inferring partner priorities) when the entire dialogue is included in the prompt. GPT-4 achieves the best performance, indicating strong out-of-the-box Theory of Mind (ToM) abilities, as opposed to smaller opensource variants that mostly perform similarly to the Majority baseline. Later in Section 5, we analyze how the model performance varies by varying the number of utterances seen by the model.

Meanwhile, for inferring the partner’s subjective assessments (Figure 3f), Flan-T5 achieves the best accuracy of 50.5%, while GPT-4 gets the highest PCC score of 0.39 – indicating that all models perform poorly when it comes to assessing the subjective perception of their partners. Prior work suggests that, whenever available, including the partner’s demographics, personality, and emotion

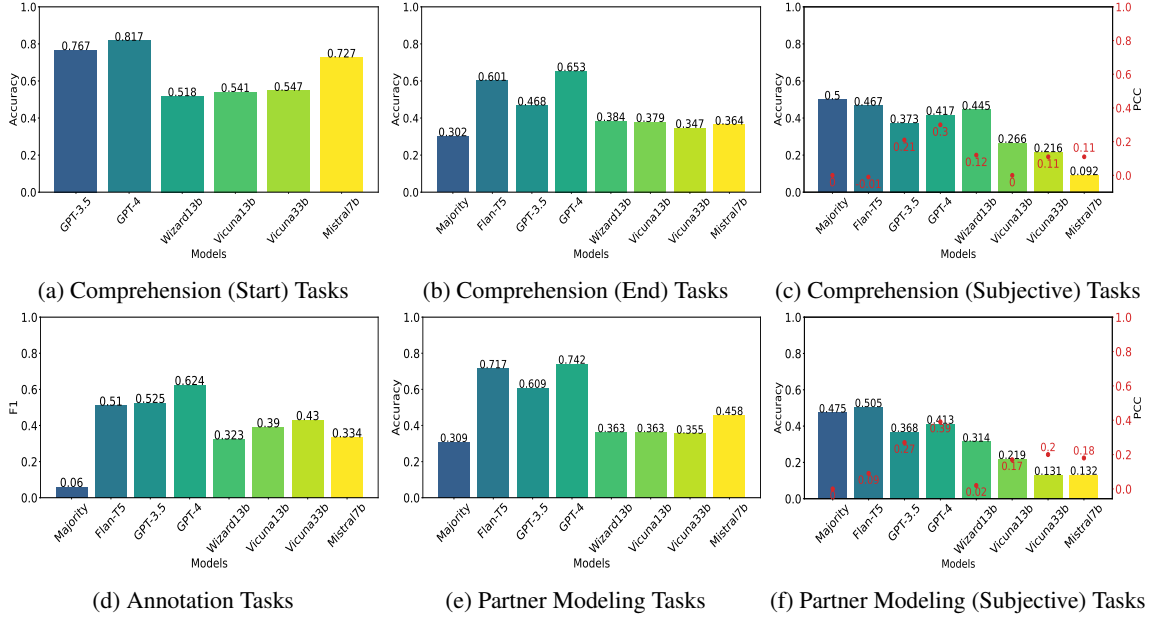


Figure 3: Overall results for zero-shot evaluation of LLMs. F1: macro F1 over all labels, PCC: Pearson Correlation Coefficient. Each bar shows the average result across all suitable tasks in the category. For example, as per (b), GPT-4 gets 65.3% Accuracy on average for *Comprehension* tasks in *End* time stage. Selected tasks for building these plots and full results are in Appendix A.

expression in the dialogue can improve these predictions (Chawla et al., 2023a).

4.4 Generation Tasks

We report the results in Table 2. Flan-T5 beats other models on BLEU and ROUGE by a huge margin, probably due to dataset-specific fine-tuning. LLMs fail to follow the utterance style and structure in the dataset, even though asked to do so in the prompts, resulting in poor overlap with the reference (examples in Appendix E). This can be attributed to the instruction-tuning step for training the LLMs.

A poor overlap with the reference does not directly imply that the response is invalid. Hence, we additionally perform an expert human evaluation (by an author of this work) on a subset of 50 examples. We measure two dimensions on a scale of 1-5: *Coherency* (How appropriate is the response given the dialogue history?) and *Strategy* (How strategic is the response given the agent context and the goal of maximizing the performance?).

GPT-4 scores the highest on *Coherency* while the Human reference baseline performs the best on *Strategy* for both datasets. However, the differences between these methods are insignificant, indicating that GPT-4 performs similarly to the average crowdsourcing worker. Further, the models perform better on *Coherency* than *Strategy* – generating a fluent and contextually appropriate

response is easier than generating a strategic response aligned with the task goals. This can be attributed to the standard autoregressive training followed for these LLMs. While the models learn to output a plausible next utterance, they fail to do so strategically – indicating task-specific tuning might still be necessary for improving performance in strategic social influence settings. Finally, the models perform much worse on *Strategy* for CA than DND, with the former involving a more complex negotiation scenario and richer language use.

GPT-4 performing similarly to the crowdsourcing workers is, although remarkable, *insufficient* for downstream applications in conversational AI and pedagogy. Prior work in negotiation agents (Lewis et al., 2017; He et al., 2018; Chawla et al., 2023d) also relies on a combination of supervised learning (learning the average behavior) and RL (explicit training for a task-specific reward). Hence, we perform an error analysis and discuss the key limitations below to guide future work.

Sophisticated Incoherent Generations: Handling off-topic interaction remains challenging, and trivial issues like repetition still persist. In some cases, we also found sophisticated coherency errors. Given the dialogue history from DND, “...*YOU: i cant accept that . how about i get the book and I hat , and you can get the balls and I hat ? THEM: i can’t accept that deal... can i have one of each ?*”,

Model	DND				CA			
	BLEU↑	Rouge-L↑	Coherency↑	Strategy↑	BLEU↑	Rouge-L↑	Coherency↑	Strategy↑
Human			4.32	4.26			4.44	3.56
Flan-T5	.167	.453	4.16	4.14	.028	.165	3.50*	2.94*
Mistral7b	.010	.130	3.86*	3.24*	.010	.130	3.84*	3.16
Wizard13b	.032	.190	3.12*	3.28*	.017	.135	3.60*	3.18
Vicuna13b	.022	.172	3.56*	3.36*	.015	.135	3.98	3.34
Vicuna33b	.038	.216	3.92	3.84	.016	.147	4.52	3.44
GPT-3.5	.030	.200	3.72*	3.60*	.025	.162	4.20	3.34
GPT-4	.017	.178	4.48	4.00	.011	.149	4.64	3.52

Table 2: Results on response generation. BLEU and Rouge-L are computed on the full testset (200 instances). Coherency and Strategy are based on a human evaluation of a random subset of 50 examples on a scale of 1-5. * means significantly worse performance than the Human reference according to a Paired Bootstrap test.

Wizard13b replies: “*YOU: I understand you want one of each. How about I get the book and one hat, and you get the balls and one hat? That way, we both get something we want.*”. Even though the output seems novel and valid, a closer look reveals that it is simply a paraphrase of YOU’s previous offer already rejected by the partner (THEM). We conclude ensuring coherency remains an open challenge and cannot be assumed to hold for LLMs.

Failure to Incorporate Strategic Cues: Generating strategically sound responses is difficult for all models. The models often fail to cater to their own priorities and struggle to utilize the information shared by the user. In one case from CA, even though the partner says, “*I am good at making fire and have a starter so I can give up all the wood.*”, the models fail to use this information in their counter offers.

Overly Agreeable Behavior: In general, the models tend to show agreement even with unfair offers, indicating that they are open for exploitation. We tried varying the prompts to adjust the model behavior, but this proved to be challenging. In the future, we suggest augmenting LLMs with task-specific RL policies for enhanced strategic reasoning.

5 Discussion

We now analyze the impact of popular prompting strategies, focusing primarily on GPT-4, given its superior overall performance.

CoT Prompting: We focus on tasks that require multi-hop arithmetic reasoning. These involve computing the maximum points possible out of a negotiation and inferring the actual total points scored in the given dialogue. We observe extremely high gains using CoT with GPT-4, achieving nearly 100% accuracy on 3 out of 4 tasks (Figure 5 of Appendix D). This provides further evidence that

CoT prompting is useful for LLMs (Madaan et al., 2023), highlighting its effectiveness in negotiation tasks as well. We provide a sample CoT prompt in Table 10 of Appendix B.

Using Prior Utterances for Annotation Tasks:

We include two prior utterances in the prompts for annotation tasks. The results are mixed (Figure 4 in Appendix D) – the performance improves in only one task while showing minor degradation in the remaining three. We expect the type of annotation labels and the data collection methodology to play a role here, but the results also indicate that irrelevant utterances might even confuse the models. Ultimately, including prior utterances must be a domain-specific choice based on the observed validation performance.

Few-shot Prompting: One of the most remarkable abilities of LLMs is In-Context Learning or ICL (Wei et al., 2022), which enables LLMs to grasp the underlying task better and follow instructions more effectively. We investigate whether adding two randomly sampled examples in annotation tasks could enhance LLMs’ performance via ICL. We find that ICL improves performance on two out of four tasks, showing evidence that few-shot examples can improve the model performance (Figure 4 in Appendix D). Selecting the best examples for ICL is an active area of research, and we expect these advancements to help tasks designed for strategic interactions as well.

Varying the number of seen utterances: We ask two research questions: 1) Do additional utterances confuse the model about its own context, which is explicitly provided in the prompt? and 2) Does the model successfully utilize the additional utterances to infer the partner’s context?

We observe that the answer to both of these questions is ‘Yes’ (Figure 6 in Appendix D). In general,

when asked about its own preferences, the models’ performance degrades when more and more utterances are supplied. However, when asked about the partner preferences, the models’ performance increases as more utterances are seen. One way to explain this is through *recency bias* – the models focus more on the recently provided information, which improves partner modeling but hurts the models’ comprehension of their own context given at the start of the input.

6 Related Work

Negotiation Agents: Lewis et al. (2017) pioneered the development of end-to-end negotiation dialogue systems using self-play Reinforcement Learning (RL). Chawla et al. (2023d) enhanced this work by employing tougher user simulators and utility-based rewards. Other efforts focused on dialogue-act based RL policies for modular agents, such as for buyer-seller negotiations (He et al., 2018). Fu et al. (2023) designed an LLM-based agent for balloon price negotiations via self-play and feedback from a critic-LLM. Gandhi et al. (2023) used LLMs to improve strategic reasoning based on few-shot CoT prompting and demonstrations about states, values, and beliefs. Abdelnabi et al. (2023) focused on interactive multi-agent games, showing that agents can consistently reach successful deals through systematic zero-shot CoT prompting. While these studies employed LLMs for designing end-to-end agents, a systematic exploration of their diverse abilities is missing, inhibiting their use in negotiation-related use cases. Hence, we evaluate LLMs across various tasks to test their multifaceted abilities in negotiation interactions.

Probing LLMs: Numerous recent efforts focus on probing LLMs’ abilities in a variety of domains (Brown et al., 2020; Kosinski, 2023; Kojima et al., 2023; Noever and McKee, 2023; Ziems et al., 2023). LLMs tend to struggle in planning and solving complex mathematical, logical, and reasoning problems (Hao et al., 2023; Huang and Chang, 2023). While they show promise in ToM tasks (Bubeck et al., 2023; Kosinski, 2023), Kosinski (2023) argue that their understanding is superficial, and Ullman (2023) provide evidence for their brittle performance. We contribute to this line of work by specifically probing LLMs in complex negotiation scenarios, covering a range of tasks from those requiring one-step reasoning to response generation that requires a blend of conversational understand-

ing, inference about the other party’s needs, and reasoning for strategic decision-making.

7 Conclusion

We devised a methodology to systematically analyze the multifaceted capabilities of LLMs in negotiations. When evaluated out-of-the-box, GPT-4 outperforms other LLMs while still leaving room for improvement for most tasks. However, CoT and few-shot prompting can somewhat fill this gap for several arithmetic reasoning and annotation tasks. In contrast, the smaller opensource models struggle to do well, performing comparably to the trivial Majority baseline for key Comprehension and Partner Modeling tasks.

Based on our results, we conclude that LLMs can indeed be helpful across a number of use cases in negotiation research. This is not only limited to designing dialogue systems but also includes scaling up pedagogical and data collection practices. LLMs capable of partner modeling can help to provide feedback to students who fail to elicit and incorporate their partner’s preferences (Johnson et al., 2019). Our results on annotation tasks indicate that LLMs can make the annotation process efficient (albeit with a human in the loop), aiding both linguistic strategy analysis and the design of modular dialogue systems. However, we note that this only holds true for proprietary LLMs, stressing the importance of continued efforts to improve smaller open-source models.

We also find that extra information in the dialogue can often confuse the models about their own context (provided explicitly in the prompts) – future work must devise ways to enable LLMs to handle longer contexts better, learning to focus only on the relevant parts of the input. Our human evaluation shows that all models struggle with generating strategically appropriate responses. Recent efforts in structured reasoning with LLMs can help to mitigate this issue (Zhou et al., 2024). These approaches are naturally suited for strategic dialogue, where response generation can be broken down into subgoals that are more feasible for the models as per our evaluations in this work, such as sequentially performing comprehension, annotation, and partner modeling before outputting the final response. In the future, we also plan to combine LLMs with RL policies explored in prior negotiation work (He et al., 2018) to control the strategy or the personality of the dialogue agent.

8 Broader Impact and Ethical Considerations

8.1 Datasets

Our study used four publicly available negotiation datasets (i.e., CRA, DND, CA, and JI), which were thoroughly anonymized before their release by the respective authors. We conducted a meticulous review of the licensing details for each dataset to ensure that our usage strictly adheres to their intended purposes and scope. We note that all datasets are in English, so it is unclear if the same findings extend to other languages or cultures. In fact, differences in how people negotiate across cultures have received significant attention in the literature (Luo, 2008; Andersen et al., 2018), and thus, we encourage future work to investigate LLM negotiation capabilities in other languages as well. Our methodology for designing tasks and evaluation procedures is language-independent, and we hope that it can guide future efforts in this direction.

8.2 LLMs

We used LLMs strictly within the intended scope in accordance with the respective licensing details. Our approach is consistent with various other recent efforts that aim to evaluate the diverse capabilities of LLMs, ensuring that the use remains within ethical and operational guidelines.

8.3 Human Evaluation

We gathered expert human annotations to evaluate the responses generated by our compared models. These were performed by the authors of this work and thus, did not require any external crowdsourcing effort.

8.4 AI for Social Influence Interactions

Negotiation dialogues fall under the broader spectrum of social influence tasks (Chawla et al., 2023c), which target achieving specific changes in behaviors or opinions through conversations (other example scenarios include online toxicity moderation, therapy, argumentation, etc.). Automated systems that can comprehend or participate in such interactions find broad applications in conversational AI and pedagogy through the development of tools that can make everyday social interactions more effective and efficient.

Ethical Recommendations: Naturally, as for any human-facing technology, efforts in this area also

raise ethical concerns that must be properly addressed. This includes possibilities for manipulation, potential misuse, bias, and discrimination (Lewicki et al., 2016).

We provide four key recommendations here: 1) Maintaining *transparency* about the dataset and model design processes, along with the known capabilities and misbehaviors of the developed systems, 2) Ensuring proper *consent* procedures, 3) Continuous *monitoring* of the designed systems, and 4) Using forums like the *ACL Workshop Series on Social Influence in Conversations (SICon)² for a *principled discussion* on this topic.

9 Limitations

Task Design: The datasets used in our analysis are based on a framework from the negotiation literature, referred to as the Multi-Issue Bargaining Task or MIBT (Fershtman, 1990). MIBT has been a popular framework for defining negotiation scenarios, both in academic and industrial settings. However, being an abstraction of real-world negotiations, it misses out on several real-world aspects, such as when the player preferences change during the interaction or when individual items can be broken down into subparts for fractional division between the players. We encourage future work to take up these other complex scenarios as well.

Prompting Variations: We primarily evaluated LLMs using zero-shot prompts to test out-of-the-box capabilities. We explored CoT and few-shot prompting for a subset of tasks to gain additional insights. Although we designed the prompts based on careful experimentation and consideration following the best practices from prior work, we acknowledge that other ways of prompting the models with more sophisticated prompt engineering methods could potentially lead to different results. This is an active area of research. While our goal in this work was to cover the breadth of capabilities based on standard prompting techniques, we encourage future work to investigate the impact of prompt engineering in-depth, albeit on a smaller number of tasks.

References

Sahar Abdelnabi, Amr Gomaa, Sarath Sivaprasad, Lea Schönherr, and Mario Fritz. 2023. [Llm-deliberation:](#)

²<https://sites.google.com/view/sicon-2023/home>

721	Evaluating llms with interactive multi-agent negotiation games.	
722		
723	Steffen Andersen, Seda Ertac, Uri Gneezy, John A List, and Sandra Maximiano. 2018. On the cultural basis of gender differences in negotiation. <i>Experimental Economics</i> , 21:757–778.	
724		
725		
726		
727	Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners.	
728		
729		
730		
731		
732		
733		
734		
735		
736		
737		
738	Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, Harsha Nori, Hamid Palangi, Marco Tulio Ribeiro, and Yi Zhang. 2023. Sparks of artificial general intelligence: Early experiments with gpt-4.	
739		
740		
741		
742		
743		
744	Kushal Chawla, Rene Clever, Jaysa Ramirez, Gale Lucas, and Jonathan Gratch. 2021a. Towards emotion-aware agents for negotiation dialogues. In <i>2021 9th International Conference on Affective Computing and Intelligent Interaction (ACII)</i> , pages 1–8. IEEE.	
745		
746		
747		
748		
749	Kushal Chawla, Rene Clever, Jaysa Ramirez, Gale M Lucas, and Jonathan Gratch. 2023a. Towards emotion-aware agents for improved user satisfaction and partner perception in negotiation dialogues. <i>IEEE Transactions on Affective Computing</i> .	
750		
751		
752		
753		
754	Kushal Chawla, Jaysa Ramirez, Rene Clever, Gale Lucas, Jonathan May, and Jonathan Gratch. 2021b. Casino: A corpus of campsite negotiation dialogues for automatic negotiation systems.	
755		
756		
757		
758	Kushal Chawla, Jaysa Ramirez, Rene Clever, Gale Lucas, Jonathan May, and Jonathan Gratch. 2021c. CaSiNo: A corpus of campsite negotiation dialogues for automatic negotiation systems. In <i>Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies</i> , pages 3167–3185, Online. Association for Computational Linguistics.	
759		
760		
761		
762		
763		
764		
765		
766	Kushal Chawla, Weiyan Shi, Jingwen Zhang, Gale Lucas, Zhou Yu, and Jonathan Gratch. 2023b. Social influence dialogue systems: A survey of datasets and models for social influence tasks.	
767		
768		
769		
770	Kushal Chawla, Weiyan Shi, Jingwen Zhang, Gale Lucas, Zhou Yu, and Jonathan Gratch. 2023c. Social influence dialogue systems: A survey of datasets and models for social influence tasks. In <i>Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics</i> , pages 750–766, Dubrovnik, Croatia. Association for Computational Linguistics.	
771		
772		
773		
774		
775		
776		
777		
	Kushal Chawla, Ian Wu, Yu Rong, Gale Lucas, and Jonathan Gratch. 2023d. Be selfish, but wisely: Investigating the impact of agent personality in mixed-motive human-agent interactions. In <i>Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing</i> , pages 13078–13092.	778 779 780 781 782 783
	Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality.	784 785 786 787 788 789
	Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling instruction-finetuned language models.	790 791 792 793 794 795 796 797 798 799 800 801
	David DeVault, Johnathan Mell, and Jonathan Gratch. 2015. Toward natural turn-taking in a virtual human negotiation agent. In <i>2015 AAAI Spring Symposium Series</i> .	802 803 804 805
	Chaim Fershtman. 1990. The importance of the agenda in bargaining. <i>Games and Economic Behavior</i> , 2(3):224–238.	806 807 808
	Chris Frith and Uta Frith. 2005. Theory of mind. <i>Current biology</i> , 15(17):R644–R645.	809 810
	Yao Fu, Hao Peng, Tushar Khot, and Mirella Lapata. 2023. Improving language model negotiation with self-play and in-context learning from ai feedback.	811 812 813
	Kanishk Gandhi, Dorsa Sadigh, and Noah D. Goodman. 2023. Strategic reasoning with language models.	814 815
	Shibo Hao, Yi Gu, Haodi Ma, Joshua Jiahua Hong, Zhen Wang, Daisy Zhe Wang, and Zhiting Hu. 2023. Reasoning with language model is planning with world model.	816 817 818 819
	He He, Derek Chen, Anusha Balakrishnan, and Percy Liang. 2018. Decoupling strategy and generation in negotiation dialogues. In <i>Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing</i> , pages 2333–2343.	820 821 822 823 824
	Jie Huang and Kevin Chen-Chuan Chang. 2023. Towards reasoning in large language models: A survey.	825 826
	Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, L�lio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timoth�e Lacroix, and William El Sayed. 2023. Mistral 7b.	827 828 829 830 831 832 833

834	Emmanuel Johnson, Gale Lucas, Peter Kim, and Jonathan Gratch. 2019. Intelligent tutoring system for negotiation skills training. In <i>International Conference on Artificial Intelligence in Education</i> , pages 122–127. Springer.	Hongru Wang, Lingzhi Wang, Yiming Du, Liang Chen, Jingyan Zhou, Yufei Wang, and Kam-Fai Wong. 2023. A survey of the evolution of language model-based dialogue systems. <i>arXiv preprint arXiv:2311.16789</i> .	885
835			886
836			887
837			888
838			
839	Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2023. Large language models are zero-shot reasoners.	Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. 2022. Emergent abilities of large language models.	889
840			890
841			891
842	Michal Kosinski. 2023. Theory of mind might have spontaneously emerged in large language models.		892
843			893
844	Roy J Lewicki, Bruce Barry, and David M Saunders. 2016. <i>Essentials of negotiation</i> . McGraw-Hill New York.	Jason Wei, Xuezi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-thought prompting elicits reasoning in large language models.	894
845			895
846			896
847	Mike Lewis, Denis Yarats, Yann N. Dauphin, Devi Parikh, and Dhruv Batra. 2017. Deal or no deal? end-to-end learning for negotiation dialogues.	Can Xu, Qingfeng Sun, Kai Zheng, Xiubo Geng, Pu Zhao, Jiazhan Feng, Chongyang Tao, and Daxin Jiang. 2023. Wizardlm: Empowering large language models to follow complex instructions. <i>arXiv preprint arXiv:2304.12244</i> .	897
848			898
849			899
850	Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization.		900
851			901
852	Peng Luo. 2008. Analysis of cultural differences between west and east in international business negotiation. <i>International Journal of Business and Management</i> , 3(11):103–106.	Atsuki Yamaguchi, Kosui Iwasa, and Katsuhide Fujita. 2021. Dialogue act-based breakdown detection in negotiation dialogues. In <i>Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume</i> , pages 745–757, Online. Association for Computational Linguistics.	902
853			903
854			904
855			905
856	Aman Madaan, Katherine Hermann, and Amir Yazdanbakhsh. 2023. What makes chain-of-thought prompting effective? a counterfactual study. In <i>Findings of the Association for Computational Linguistics: EMNLP 2023</i> , pages 1448–1535.	Pei Zhou, Jay Pujara, Xiang Ren, Xinyun Chen, Heng-Tze Cheng, Quoc V Le, Ed H Chi, Denny Zhou, Swaroop Mishra, and Huaixiu Steven Zheng. 2024. Self-discover: Large language models self-compose reasoning structures. <i>arXiv preprint arXiv:2402.03620</i> .	906
857			907
858			908
859			909
860			910
861	Johnathan Mell and Jonathan Gratch. 2017. Grumpy & pinocchio: answering human-agent negotiation questions through realistic agent design. In <i>Proceedings of the 16th conference on autonomous agents and multiagent systems</i> , pages 401–409.	Caleb Ziems, William Held, Omar Shaikh, Jiaao Chen, Zhehao Zhang, and Diyi Yang. 2023. Can large language models transform computational social science? <i>arXiv preprint arXiv:2305.03514</i> .	911
862			912
863			913
864			914
865			915
866	Johnathan Mell, Jonathan Gratch, Reyhan Aydoğan, Tim Baarslag, and Catholijn M Jonker. 2019. The likeability-success tradeoff: Results of the 2 nd annual human-agent automated negotiating agents competition. In <i>2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII)</i> , pages 1–7. IEEE.		916
867			917
868			918
869			919
870			
871			
872			
873	David Noever and Forrest McKee. 2023. Numeracy from literacy: Data science as an emergent skill from large language models.		
874			
875			
876	Richard L Oliver, PV Sundar Balakrishnan, and Bruce Barry. 1994. Outcome satisfaction in negotiation: A test of expectancy disconfirmation. <i>Organizational Behavior and Human Decision Processes</i> , 60(2):252–275.		
877			
878			
879			
880			
881	OpenAI. 2022. Openai: Introducing chatgpt.		
882	OpenAI. 2023. Openai: Gpt-4.		
883	Tomer Ullman. 2023. Large language models fail on trivial alterations to theory-of-mind tasks.		
884			

A Details of Negotiation Tasks

We present an overview of all the designed tasks in Table 3. Table 4 summarizes the distribution of tasks across datasets and time stages. In some cases, a specific model failed to follow the provided instructions for a specific task and hence, failed to produce reliable results. Such tasks were entirely excluded (for all models) while computing the average bar plots in Figure 3 for a fair comparison. Table 5 presents the final list of selected tasks used for computing these plots. Finally, we report the task-wise results for all models in Table 6 for completeness.

B Prompting

The precise versions used for GPT-3.5 and GPT-4 are ‘gpt-3.5-turbo-1106’ and ‘gpt-4-1106-preview’, respectively. Tables 7, 8, and 9 provide examples of prompts for specific tasks corresponding to the *start*, *during*, and *end* time stages respectively. Additionally, Table 10 showcases an example of CoT prompting used for the *end_deal_total_ca* task. Table 11 shows the questions used in prompts for all tasks.

C Baselines

We use two baselines in this work: a *Majority* baseline, that trivially outputs the majority answer from the training data, and a *Flan-T5* fine-tuning baseline, where we train a model for each task separately by fine-tuning on the available training data. **Task Selection:** For all tasks that are defined in the *Start* time stage, it is possible to achieve a 100% accuracy through trivial rule-based parsers. Hence, we excluded these tasks for evaluation with *Majority* and *Flan-T5*.

Data Preprocessing: During the data preprocessing for fine-tuning, although there are differences in the information included in the prompts for each task, prompts for evaluating LLMs are generally lengthy due to detailed task descriptions. To enhance training efficiency during fine-tuning, we removed these lengthy descriptions from the inputs to the Flan-T5 model and instead replaced them with simple task instructions such as ‘predict highest priority’, similar to using brief descriptions when typically pretraining Flan-T5 in a multi-task setting.

After applying the previously described preprocessing methods, we constructed the final dataset and divided it for training and evaluation purposes

into train, dev, and test sets with a ratio of 8:1:1, respectively.

Multi-task Training In line with how Flan-T5 was originally trained, we conducted our experiments in two ways: 1) Training a single model on all tasks together (FT-5-All-Task) and 2) Training one model for each individual task separately (FT-5-By-Task). The results showed that the FT-5-All-Task model generally underperformed compared to the FT-5-By-Task models, with observed poor learning in several tasks. Consequently, FT-5-By-Task models were chosen as baseline.

Compute Resources: We trained the baseline model (Flan-T5) for each task under various experimental settings and hyperparameter adjustments, utilizing over 500 hours of GPUs such as NVIDIA V100 and A100 GPUs.

Implementation: For fine-tuning Flan-T5, we used the released model³ on the Hugging Face model hub as a back-bone model. The evaluation code for model assessment was developed from scratch, while for evaluating generated responses, we utilized existing packages for BLEU⁴ and ROUGE-L⁵ respectively.

D Task-specific Analysis

Figure 4 presents the results for few-shot prompting and incorporating previous utterances for various *Annotation* tasks. Figure 5 shows the results for Chain-of-Thought (CoT) prompting technique on four multi-hop arithmetic reasoning *Comprehension* tasks. Finally, we present the effect of varying the number of utterances seen by the models for preference prediction tasks in Figure 6.

E Generation Outputs

We present sample generations by the compared models along with the ground-truth reference in Tables 12 and 13.

³<https://huggingface.co/google/flan-t5-base>

⁴<https://pypi.org/project/evaluate/>

⁵<https://pypi.org/project/rouge-score/>

Dataset	Time Stage	Full Task Name	Task Type
CA	Start	sta_ask_high_priority_ca	Comprehension
CA	Start	sta_ask_low_priority_ca	Comprehension
CA	Start	sta_ask_point_values_ca	Comprehension
CA	Start	sta_max_points_ca	Comprehension
CA	Start	sta_total_item_count_ca	Comprehension
CA	During	dur_partner_ask_high_priority_ca	Partner Modeling
CA	During	dur_partner_ask_low_priority_ca	Partner Modeling
CA	During	dur_strategy_ca	Annotation
CA	During	dur_gen_resp_ca	Generation
CA	During	dur_ask_high_priority_ca	Comprehension
CA	During	dur_ask_low_priority_ca	Comprehension
CA	End	end_deal_likeness_ca	Comprehension
CA	End	end_deal_satisfaction_ca	Comprehension
CA	End	end_deal_specifics_ca	Comprehension
CA	End	end_deal_total_ca	Comprehension
CA	End	end_partner_deal_likeness_ca	Partner Modeling
CA	End	end_partner_deal_satisfaction_ca	Partner Modeling
CRA	During	dur_dial_act_cra	Annotation
CRA	During	dur_full_proposal_cra	Annotation
DND	Start	sta_ask_point_values_dnd	Comprehension
DND	Start	sta_max_points_dnd	Comprehension
DND	Start	sta_total_item_count_dnd	Comprehension
DND	During	dur_dial_act_dnd	Annotation
DND	During	dur_full_proposal_dnd	Annotation
DND	During	dur_gen_resp_dnd	Generation
DND	End	end_deal_specifics_dnd	Comprehension
DND	End	end_deal_total_dnd	Comprehension
JI	Start	sta_ask_high_priority_ji	Comprehension
JI	Start	sta_ask_low_priority_ji	Comprehension
JI	During	dur_dial_act_ji	Annotation
JI	During	dur_partner_ask_high_priority_ji	Comprehension
JI	During	dur_partner_ask_low_priority_ji	Comprehension
JI	During	dur_ask_high_priority_ji	Comprehension
JI	During	dur_ask_low_priority_ji	Comprehension
JI	End	end_deal_specifics_ji	Comprehension

Table 3: An overview of the designed tasks for evaluating LLMs in negotiations. *CRA*: (DeVault et al., 2015), *DND*: (Lewis et al., 2017), *CA*: (Chawla et al., 2021c), *JI*: (Yamaguchi et al., 2021).

Dataset	Negotiation Stage			Total
	Start	During	End	
CA	5	6	6	17
CRA		2		2
DND	3	3	2	8
JI	2	5	1	8
Total	10	16	9	35

Table 4: Distribution of the designed tasks by the dataset and time stage.

Task Types	Task Names
Comprehension (Start)	sta_max_points_ca, sta_max_points_dnd, sta_total_item_count_ca, sta_total_item_count_dnd, sta_ask_high_priority_ji, sta_ask_low_priority_ji
Comprehension (End)	end_deal_specifics_ca, end_deal_specifics_dnd, end_deal_total_ca, end_deal_total_dnd
Comprehension (Subjective)	end_deal_satisfaction_ca
Annotation (During)	dur_dial_act_cra, dur_dial_act_ji, dur_strategy_ca
Partner Modeling (During)	dur_partner_ask_high_priority_ca, dur_partner_ask_low_priority_ca
Partner Modeling (Subjective)	end_deal_satisfaction_ca, end_deal_likeness_ca

Table 5: Selected tasks for computing the average bar plots in Figure 3.

Full Task Name	Metric	Model							
		Majority	Flan-T5	GPT-3.5	GPT-4	Mistral7b	Vicuna13b	Vicuna33b	Wizard13b
end_deal_likeness_ca	PCC	0.525	0.525	0.357	0.175		0.119	0.267	0.239
end_deal_satisfaction_ca	PCC	0.5	0.467	0.373	0.417	0.092	0.266	0.216	0.445
end_deal_specifics_ca	Accuracy	0.356	0.364	0.664	0.916	0.517	0.517	0.593	0.555
end_deal_total_ca	Accuracy	0.142	0.233	0.158	0.083	0.15	0.05	0.017	0.017
end_partner_deal_likeness_ca	PCC	0.517	0.517	0.31	0.308	0.133	0.167	0.178	0.282
end_partner_deal_satisfaction_ca	PCC	0.433	0.492	0.426	0.517	0.13	0.271	0.083	0.345
dur_ask_high_priority_ca	Accuracy			0.742	0.9	0.558		0.375	0.345
dur_ask_low_priority_ca	Accuracy			0.533	0.75	0.358		0.286	0.269
dur_partner_ask_high_priority_ca	Accuracy	0.292	0.717	0.7	0.792	0.483	0.42	0.353	0.392
dur_partner_ask_low_priority_ca	Accuracy	0.325	0.717	0.517	0.692	0.433	0.306	0.357	0.333
dur_strategy_ca	F1	0.055	0.724	0.463	0.507	0.265	0.381	0.304	0.254
sta_ask_high_priority_ca	Accuracy			1	1				0.667
sta_ask_low_priority_ca	Accuracy			1	1	0.5			0.4
sta_ask_point_values_ca	F1			1	1	1	1	1	1
sta_max_points_ca	Accuracy			0.333	0.333	0.5	0	0	0
sta_total_item_count_ca	Accuracy			1	1	1	1	1	0.333
dur_dial_act_cra	F1	0.067	0.787	0.535	0.678	0.35	0.338	0.518	0.302
dur_full_proposal_cra	Accuracy	0.359	0.439	0.352	0.369	0.241	0.262	0.245	0.325
end_deal_specifics_dnd	Accuracy	0.454	0.973	0.67	0.949	0.558	0.631	0.558	0.628
end_deal_total_dnd	Accuracy	0.257	0.832	0.381	0.664	0.23	0.319	0.221	0.336
dur_dial_act_dnd	F1	0.888	0.96	0.735	0.825	0.764		0.639	0.337
dur_full_proposal_dnd	Accuracy	0.39	1	0.742	0.866	0.648	0.748	0.725	0.687
sta_ask_point_values_dnd	Accuracy			0.993	1	1	1	0.752	1
sta_max_points_dnd	Accuracy			0.317	0.337	0.366	0.495	0.307	0.386
sta_total_item_count_dnd	Accuracy			0.95	1	0.98	0.505	0.901	0.465
end_deal_specifics_ji	Accuracy	0.261	0.764	0.782	0.858	0.733	0.8	0.785	0.766
dur_ask_high_priority_ji	Accuracy			0.495	0.862	0.37	0.233	0.252	0.259
dur_ask_low_priority_ji	Accuracy			0.67	0.917	0.333	0.26	0.306	0.296
dur_dial_act_ji	F1	0.058	0.019	0.578	0.688	0.387	0.452	0.468	0.414
dur_partner_ask_high_priority_ji	Accuracy	0.165	0.202	0.193			0.198	0.204	0.204
dur_partner_ask_low_priority_ji	Accuracy	0.193	0.266	0.202		0.269	0.176	0.157	0.13
sta_ask_high_priority_ji	Accuracy			0.78	0.89	0.505	0.155	0.211	0.596
sta_ask_low_priority_ji	Accuracy			0.761	0.972	0.468	0.174	0.202	0.367

Table 6: Task-wise results for all models. Empty values for Majority and Flan-T5 correspond to the tasks on which these baselines were not evaluated (since a simple rule-based baseline can achieve 100% performance in these cases). Empty values in all other cases correspond to scenarios where the LLM model failed to follow task-specific instructions, making the measured scores too unreliable to report. Notes: 1) PCC: Pearson Correlation Coefficient, 2) The results on response generation are provided in Table 2, and 3) For the cases where the models guess the outputs for each issue (like books, balls, or hats in DND) separately, we simply report the average score across all issues. The results for two generation tasks (i.e., *dur_gen_resp_ca*, *dur_gen_resp_dnd*) are excluded from the table and can be seen in Table 2.

Prompt example (Task: sta_ask_point_values_ca)

Task Description: You are negotiating with your campsite neighbor over an extra supply of food, water, and firewood for your camping trip. Different types of packages are worth different amounts of points to each one of you. You'll be provided with information about the negotiation. Then, you'll answer a question.

Here are the number of food, water, and firewood packages available in the negotiation, contained in <count>tags.
<count>Food Packages: 3 Water Packages: 3 Firewood Packages: 3 </count>

Here are the number of points you get for each type of package, contained in <value>tags.
<value>Each Food Package: 3 points Each Water Package: 5 points Each Firewood Package: 4 points </value>

Question: How many points is one package of each issue worth to you? Present your answer as a json within <answer></answer>tags with keys as issues (food, water, and firewood) and values as the corresponding answers.

Table 7: Prompt example for the *sta_ask_point_values_ca* task.

Prompt example (Task: dur_full_proposal_dnd)

Task Description: You are negotiating with a partner over some quantity of books, hats, and balls to determine who gets which items. Different types of items are worth different amount of points to each one of you. You'll be provided with information about the negotiation. Then, you'll answer a question.

Here are the number of books, hats, and balls available in the negotiation, contained in <count>tags.
<count>Books: 3 Hats: 1 Balls: 2 </count>

Here are the number of points you get for each type of item, contained in <value>tags.
<value>Each Book: 1 points Each Hat: 5 points Each Ball: 1 points </value>

Here is an utterance from the negotiation, contained in <utterance>tags.
<utterance>YOU: i'll take the hat and balls if you want the books </utterance>

Question: How many items does the speaker get for each issue in the proposal delimited by the <utterance>tags? Present your answer as a json within <answer></answer>tags with keys as issues (books, hats, and balls) and values as the corresponding answers. If the answer is not clear for an issue, output NA.

Table 8: Prompt example for the *dur_full_proposal_dnd* task.

Prompt example (Task: end_deal_specifics_ca)

Task Description: You are negotiating with your campsite neighbor over extra supply of food, water, and firewood for your camping trip. Different types of packages are worth different amount of points to each one of you. You'll be provided with information about the negotiation. Then, you'll answer a question.

Here are the number of food, water, and firewood packages available in the negotiation, contained in <count>tags.
<count>Food Packages: 3 Water Packages: 3 Firewood Packages: 3 </count>

Here are the number of points you get for each type of package, contained in <value>tags.
<value>Each Food Package: 3 points Each Water Package: 5 points Each Firewood Package: 4 points </value>

Here is the complete dialogue, contained in <dialogue>tags.
<dialogue>THEM: Hello, I would like to have three packages of food. We've decided to stay an extra night but need more food to do so.
YOU: I would be open to that if you could give me three packages of water ☺
THEM: Hmm...I'm pretty muddy due to clumsiness, so I may need one extra. I could give you two waters and all of the firewood. What do you think? ☺
YOU: So are you suggesting that I would get 2 waters, 3 firewood, and no food?
THEM: Right! Well, beyond the food you already have.
YOU: I have an extra person camping with us that I didn't expect when I bought food, so I could use one if you're willing ☺
THEM: I understand that! I wasn't expecting to stay an extra night, but the weather is too perfect to leave. I can manage with two packages of food for sure. ☺
YOU: Great! Thank you for being so understanding!
THEM: No problem! So are we in agreement that I get 2 food, 1 water and you get the reverse? I could also probably use one firewood, but it's not as important to me.
YOU: I can give you one firewood, so I'll be getting 1 food, 2 water, and 2 firewood? </dialogue>

Question: In the final deal, how many item of each issue did you get? Present your answer as a json within <answer></answer>tags with keys as issues (food, water, and firewood) and values as the corresponding answers. If there was no agreement, answer NA for each issue.

Table 9: Prompt example for the *end_deal_specifics_ca* task.

Prompt example (Task: end_deal_total_ca)

Task Description: You are negotiating with your campsite neighbor over extra supply of food, water, and firewood for your camping trip. Different types of packages are worth different amount of points to each one of you. You'll be provided with information about the negotiation. Then, you'll answer a question.

Here are the number of food, water, and firewood packages available in the negotiation, contained in <count> tags.

<count>
 Food Packages: 3
 Water Packages: 3
 Firewood Packages: 3
 </count>

Here are the number of points you get for each type of package, contained in <value> tags.

<value>
 Each Food Package: 3 points
 Each Water Package: 5 points
 Each Firewood Package: 4 points
 </value>

Here is the complete dialogue, contained in <dialogue> tags.

<dialogue>
 THEM: Hello, I would like to have three packages of food. We've decided to stay an extra night but need more food to do so.
 YOU: I would be open to that if you could give me three packages of water
 THEM: Hmm...I'm pretty muddy due to clumsiness, so I may need one extra. I could give you two waters and all of the firewood. What do you think?
 YOU: So are you suggesting that I would get 2 waters, 3 firewood, and no food?
 THEM: Right! Well, beyond the food you already have.
 YOU: I have an extra person camping with us that I didn't expect when I bought food, so I could use one if you're willing
 THEM: I understand that! I wasn't expecting to stay an extra night, but the weather is too perfect to leave. I can manage with two packages of food for sure.
 YOU: Great! Thank you for being so understanding!
 THEM: No problem! So are we in agreement that I get 2 food, 1 water and you get the reverse? I could also probably use one firewood, but it's not as important to me.
 YOU: I can give you one firewood, so I'll be getting 1 food, 2 water, and 2 firewood?
 </dialogue>

Question: How many points did you get at the end of the negotiation?

NOTE: Let's think step-by-step! Put your thoughts in <thinking> </thinking> tags, and put your answer as a single number in <answer> </answer> tags.

Table 10: CoT prompt example for the *end_deal_total_ca* task.

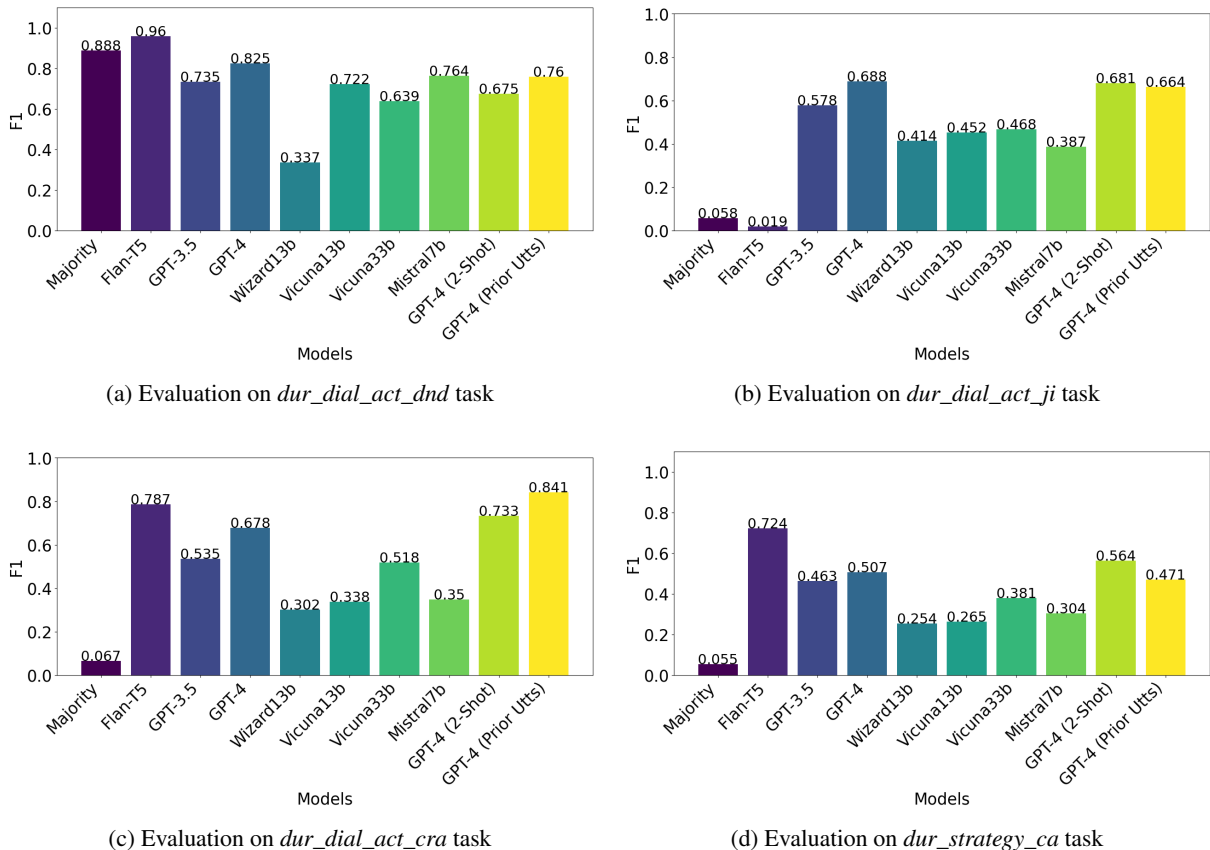


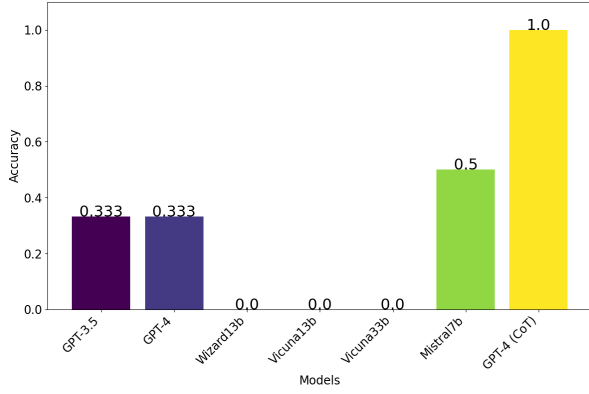
Figure 4: Evaluation on four tasks demonstrating the impact of 1) Two-shot prompting and 2) Incorporating two prior utterances. We only consider GPT-4 for this analysis. F1 refers to the macro F1 score over all labels.

Task	Question
<i>sta_total_item_count_dnd</i> <i>sta_total_item_count_ca</i>	What is the total number of items being negotiated over? Present your answer as a single number with no additional text.
<i>sta_max_points_dnd</i> <i>sta_max_points_ca</i>	What is the maximum number of points that you can possibly get in any deal? Present your answer as a single number with no additional text.
<i>sta_ask_point_values_dnd</i>	How many points is one item of each issue worth to you? Present your answer as a JSON within <code><answer></answer></code> tags with keys as issues (books, hats, and balls) and values as the corresponding answers.
<i>sta_ask_point_values_ca</i>	How many points is one package of each issue worth to you? Present your answer as a JSON within <code><answer></answer></code> tags with keys as issues (food, water, and firewood) and values as the corresponding answers.
<i>sta_ask_low_priority_ji</i> <i>dur_ask_low_priority_ji</i>	What is your lowest priority issue? Present your answer as one of the following multiple choice options. You must select an option. A: position / B: company / C: salary / D: days_off / E: workplace
<i>sta_ask_low_priority_ca</i> <i>dur_ask_low_priority_ca</i>	What is your lowest priority issue? Present your answer as one of the following multiple choice options. You must select an option. A: food / B: water / C: firewood
<i>sta_ask_high_priority_ji</i> <i>dur_ask_low_priority_ca</i>	What is your highest priority issue? Present your answer as one of the following multiple choice options. You must select an option. A: position / B: company / C: salary / D: days_off / E: workplace
<i>sta_ask_high_priority_ca</i> <i>dur_ask_high_priority_ca</i>	What is your highest priority issue? Present your answer as one of the following multiple choice options. You must select an option. A: food / B: water / C: firewood
<i>dur_strategy_ca</i>	Which negotiation strategies are employed in the utterance? Present your answer as a comma-separated list of strategies, contained in <code><answer></answer></code> tags with no additional text.
<i>dur_partner_ask_low_priority_ji</i>	What is the recruiter's lowest priority issue? Present your answer as one of the following multiple choice options. You must select an option. A: position / B: company / C: salary / D: days_off / E: workplace
<i>dur_partner_ask_low_priority_ca</i>	What is your partner's lowest priority issue? Present your answer as one of the following multiple choice options. You must select an option. A: food / B: water / C: firewood
<i>dur_partner_ask_high_priority_ji</i>	What is the recruiter's highest priority issue? Present your answer as one of the following multiple choice options. You must select an option. A: position / B: company / C: salary / D: days_off / E: workplace
<i>dur_partner_ask_high_priority_ca</i>	What is your partner's highest priority issue? Present your answer as one of the following multiple choice options. You must select an option. A: food / B: water / C: firewood
<i>dur_gen_resp_dnd</i> <i>dur_gen_resp_ca</i>	Given the recent dialogue history inside <code><dialogue></code> tags, generate your next response in the negotiation concisely, following a similar style as previous utterances.
<i>dur_full_proposal_dnd</i>	How many items does the speaker get for each issue in the proposal delimited by the <code><utterance></code> tags? Present your answer as a JSON within <code><answer></answer></code> tags with keys as issues (books, hats, and balls) and values as the corresponding answers. If the answer is not clear for an issue, pick your best guess.
<i>dur_full_proposal_ca</i>	How many items does the speaker get for each issue in the proposal delimited by the <code><utterance></code> tags? Present your answer as a JSON within <code><answer></answer></code> tags with keys as issues (painting, lamp, and record) and values as the corresponding answers. If the answer is not clear for an issue, output NA.
<i>dur_dial_act_ji</i> <i>dur_dial_act_ca</i>	Which dialogue acts are employed in the utterance delimited by the <code><utterance></code> tags? Present your answer as a Python list of the relevant options. At least one option applies.
<i>dur_dial_act_dnd</i>	Which dialogue act is employed in the utterance contained in <code><utterance></code> tags? Present your answer as a single word.
<i>end_partner_deal_satisfaction_ca</i>	How satisfied do you think your partner is with the negotiation outcome? Present your answer as one of the following multiple choice options. You must select an option. A: extremely_dissatisfied / B: slightly_dissatisfied / C: undecided / D: slightly_satisfied / E: extremely_satisfied
<i>end_partner_deal_likeness_ca</i>	How much do you think your partner likes you? Present your answer as one of the following multiple choice options. You must select an option. A: extremely_dissatisfied / B: slightly_dissatisfied / C: undecided / D: slightly_satisfied / E: extremely_satisfied
<i>end_deal_total_dnd</i> <i>end_deal_total_ca</i>	How many points did you get at the end of the negotiation? Present your answer as a single number with no additional text.
<i>end_deal_specifics_ji</i>	In the final deal, what value was agreed on for each issue? Present your answer as a JSON within <code><answer></answer></code> tags with keys as issues (Company, Position, Workplace, Salary, Days_off) and values as the corresponding answers. If you are unsure, pick your best guess.
<i>end_deal_specifics_dnd</i>	In the final deal, how many items of each issue did you get? Present your answer as a JSON within <code><answer></answer></code> tags with keys as issues (books, hats, and balls) and values as the corresponding answers. If you are unsure, pick your best guess.
<i>end_deal_specifics_ca</i>	In the final deal, how many items of each issue did you get? Present your answer as a json within <code><answer></answer></code> tags with keys as issues (food, water, and firewood) and values as the corresponding answers. If you are unsure, pick your best guess.
<i>end_deal_satisfaction_ca</i>	How satisfied are you with the negotiation outcome? Present your answer as one of the following multiple choice options. You must select an option. A: extremely_dissatisfied / B: slightly_dissatisfied / C: undecided / D: slightly_satisfied / E: extremely_satisfied
<i>end_deal_likeness_ca</i>	How much do you like your partner? Present your answer as one of the following multiple choice options. You must select an option. A: extremely_dissatisfied / B: slightly_dissatisfied / C: undecided / D: slightly_satisfied / E: extremely_satisfied

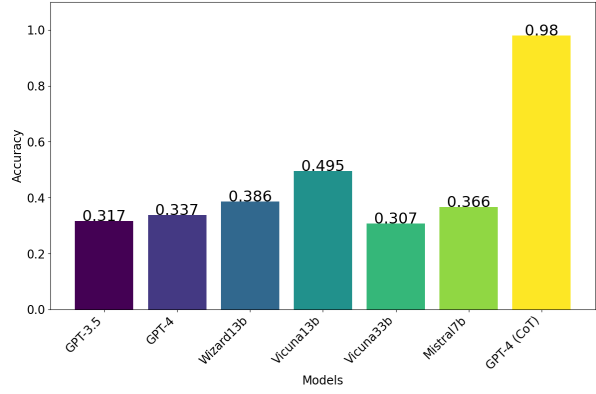
Table 11: Questions used in the prompts for all tasks.

Counts	Books: 1, Hats: 3, Balls: 2		
Agent Values	Each Book: 1 points, Each Hat: 1 points, Each Ball: 3 points		
Dialogue	THEM: i'm interested in the book and both balls in exchange for the hats . YOU: definitely not both balls . we could split them ? THEM: i think that's fair . i agree . 1 book , 1 ball for me . . . rest for you .		
	Response		
Reference	YOU: deal !	Coherency	Strategy
Flan-T5	YOU: deal.	5	5
Mistral7b	YOU: I understand your offer, but I'm not willing to give up both balls. How about we split the balls and the hats evenly? That way, we can both get what we want.	2	4
Wizard13b	YOU: 1 book, 1 ball for me.	1	2
Vicuna13b	YOU: 1 book, 1 ball for you, and the rest for me.	2	5
Vicuna33b	YOU: Sounds good. I'll take the remaining book and both hats, and you can have the ball.	2	3
GPT-3.5	YOU: Sounds good. I'll take the 3 hats then. Deal.	5	5
GPT-4	YOU: Great, so to confirm, I'll take 2 hats and 1 ball, and you get 1 book and 1 ball. Do we have a deal?	3	4

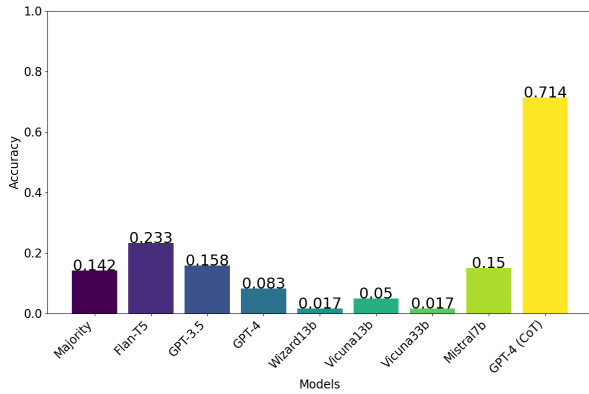
Table 12: Sample responses generated by the models for an instance from the DND dataset. The *Coherency* and *Strategy* scores are based on human annotations on a scale of 1-5.



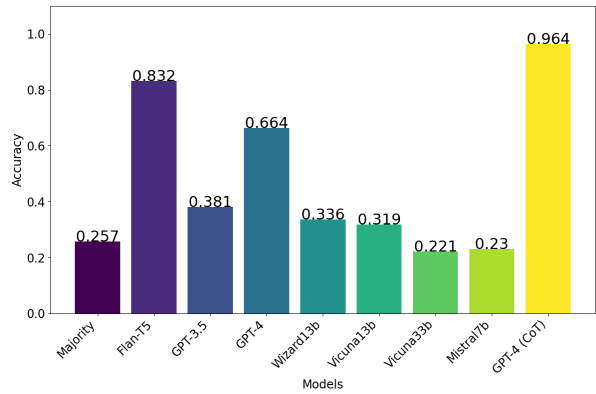
(a) Evaluation on *sta_max_point_ca* task



(b) Evaluation on *sta_max_point_dnd* task



(c) Evaluation on *end_deal_total_ca* task

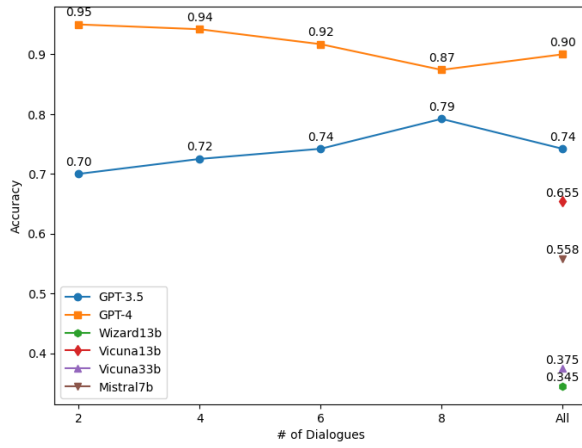


(d) Evaluation on *end_deal_total_dnd* task

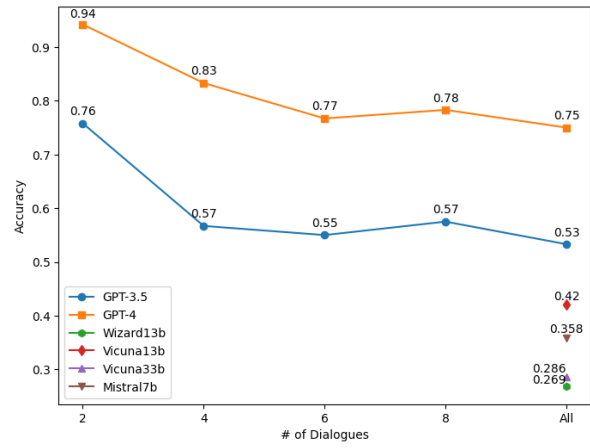
Figure 5: Evaluation on four tasks demonstrating the impact of CoT prompting. We only consider GPT-4 for this analysis.

Counts	Food Packages: 3, Water Packages: 3, Firewood Packages: 3		
Agent Values	Each Food Package: 3 points, Each Water Package: 4 points, Each Firewood Package: 5 points		
Dialogue	<p>YOU: hello</p> <p>THEM: Hello, there. Should we discuss how to divide up the items for the camping trip?</p> <p>YOU: yeah. i need 2 firewood, 1 water and 1 food</p> <p>THEM: I need to have 2 or 3 pieces of firewood because I did not pack any way to heat up my food without firewood.</p>		
	Response	Coherency	Strategy
Reference	YOU: I also need 2 pieces of firewood	4	3
Flan-T5	YOU: i need 2 firewood, 2 water and 1 food	3	3
Mistral7b	YOU: Based on the recent dialogue history, it seems like we both need firewood for our camping trip. How about we agree to split the 3 firewood packages equally, and I'll take the 1 water and 1 food package?	5	4
Wizard13b	YOU: I understand, I can provide you with 2 firewood, 1 water, and 1 food. How about we split the remaining items equally?	5	3
Vicuna13b	YOU: I understand your need for firewood. How about we split the firewood packages evenly, with each of us getting 1 and a half packages? That way, you'll have enough firewood for your needs, and I'll still have some for my camping trip.	5	3
Vicuna33b	YOU: I understand your need for firewood. Let's propose a deal: I'll take 1 firewood, 1 water, and 2 food packages. In return, you can have the remaining 2 firewood packages. This way, we both get what we need.	5	2
GPT-3.5	YOU: I can give you 2 firewood, 1 water, and 1 food package.	4	5
GPT-4	YOU: How about if I take 2 firewood and you take 1, and in exchange, I can let you have 2 water packages?	4	4

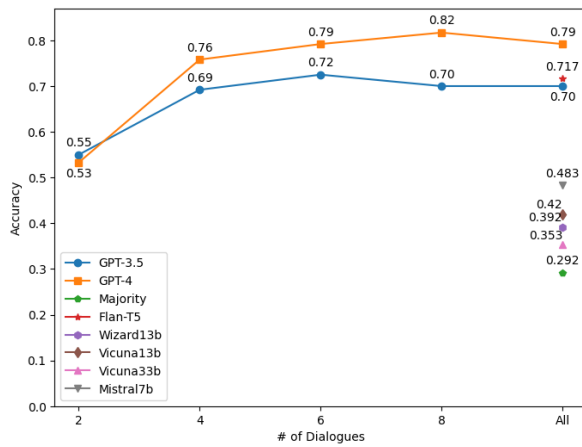
Table 13: Sample responses generated by the models for an instance from the CaSiNo dataset. The *Coherency* and *Strategy* scores are based on human annotations on a scale of 1-5.



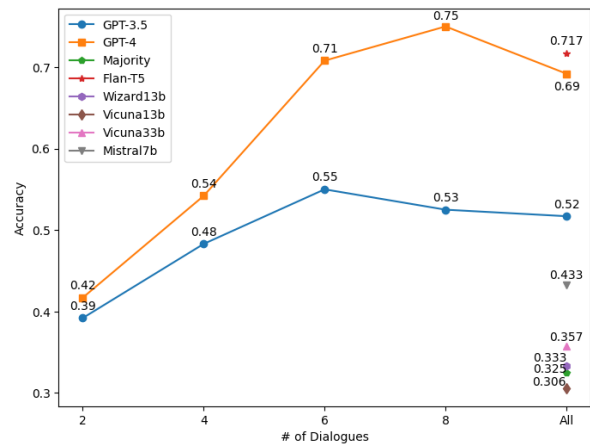
(a) Evaluation on *dur_ask_high_priority_ca* task



(b) Evaluation on *dur_ask_low_priority_ca* task



(c) Evaluation on *dur_partner_ask_high_priority_ca* task



(d) Evaluation on *dur_partner_ask_low_priority_ca* task

Figure 6: Evaluation on four tasks demonstrating the impact of increasing the number of utterances seen by the models. We only consider GPT-3.5 and GPT-4 for this analysis.