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

Deuksin Kwon[1] **Emily Weiss**[1] **Tara Kulshrestha**[2]
Kushal Chawla[3] **Gale M. Lucas**[1] **Jonathan Gratch**[1]
[1]University of Southern California [2]University of California, Berkeley
[3]Capital One
{deuksink,emilymwe}@usc.edu
tarakul@berkeley.edu kushal.chawla@capitalone.com
{lucas,gratch}@ict.usc.edu

Abstract

A successful negotiation requires a range of capabilities, including comprehension of the conversation context, Theory-of-Mind (ToM) skills to infer the partner's motives, strategic reasoning, and effective communication, making it challenging for automated systems. Despite the remarkable performance of LLMs in various NLP tasks, there is no systematic evaluation of their capabilities in negotiation. Such an evaluation is critical for advancing AI negotiation agents and negotiation research, ranging from designing dialogue systems to providing pedagogical feedback and scaling up data collection practices. This work aims to systematically analyze the multifaceted capabilities of LLMs across diverse dialogue scenarios throughout the stages of a typical negotiation interaction. Our analysis highlights GPT-4's superior performance in many tasks while identifying specific challenges, such as making subjective assessments and generating contextually appropriate, strategically advantageous responses. The code is available at https://github.com/DSincerity/SysEval-NegoLLMs

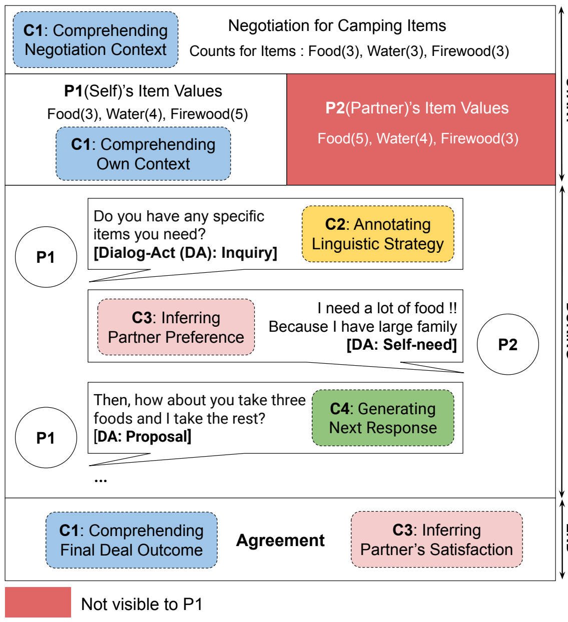


Figure 1: Overview of the key capabilities (C1-C4) required for a successful negotiation. We design tasks aligned with these abilities to assess how LLMs can advance different aspects of negotiation research. The negotiation scenario is based on Chawla et al. (2021b).

1 Introduction

Large Language Models (LLMs), known for their impressive understanding and reasoning capabilities, are increasingly being utilized as AI negotiation agents (Fu et al., 2023). However, prior research on LLM-based negotiation agents lacks a systematic analysis and evaluation of the diverse abilities required for effective negotiation (Fu et al., 2023; Gandhi et al., 2023; Abdelnabi et al., 2023). To develop more advanced AI negotiating agents, it is crucial to assess the various capabilities necessary for negotiation. Systematic evaluation of LLMs through diverse negotiation tasks can provide interpretable insights by breaking down the complex negotiation process into single tasks. This approach significantly enhances the utility and value of LLMs for various purposes in negotiation dialogues (e.g., 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)).

For a sophisticated evaluation of the multifaceted capabilities of LLMs in negotiation, it is essential to consider the core competencies required by agents in a realistic yet challenging bilateral multi-issue negotiation scenario (Fershtman, 1990), as illustrated in Figure1. Players begin with predefined value preferences for certain issues (e.g., food, water, firewood) and engage in dialogue to divide

Findings of the Association for Computational Linguistics: EMNLP 2024, pages 5391–5413
November 12-16, 2024

all available items, aiming to maximize their total value. To succeed, a negotiation agent must understand the negotiation context and outcome (**C1**), grasp the dialogue's structure (e.g., intent) and semantics (e.g., linguistic strategies) (**C2**), infer the partner's preferences through conversation (**C3**), and generate coherent, strategically apt responses given negotiation context and conversation (**C4**). These challenges highlight negotiation as a rigorous AI research testbed.

To this end, we propose a comprehensive framework that systematically analyzes state-of-the-art LLMs in negotiations across 35 tasks, considering the varied skills required by competent negotiation agents. Specifically, successful negotiation requires the Theory of Mind (ToM) (Frith and Frith, 2005)—the ability to understand the opponent's perspective—and the capacity to infer subjective aspects like affective states, beyond objective information. Our framework encompasses a range of partner modeling and subjective tasks, defined along three crucial axes: 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). Our contributions are as follows:

1. We design a novel methodology to evaluate the multifaceted capabilities of LLMs in negotiations. While our framework is specifically designed with the goal of supporting negotiation research, the resulting methodology also captures nuances common to other dialogue tasks (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 an evaluation with expert human annotators, we find that GPT-4 performs comparably to humans in response generation. We perform an error analysis, discussing the Coherence and strategic failures that still persist in the compared LLMs (Section 4.4).

4. We uncover intriguing results, finding the effectiveness of CoT and few-shot prompting. Also, by evaluating LLMs' prediction balance, we confirm GPT-4's superior robustness compared to other models (Section 5).

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 task-specific prompts, and 3) Evaluating and analyzing various proprietary and open-source LLMs.

2.1 Datasets

Three key requirements guided our dataset selections: 1) Well-structured design with concrete player preferences and outcomes, 2) Diversity in task scenarios, and 3) Availability of metadata for testing holistic LLM abilities. Based on these criteria, we selected four datasets: *CRA* (DeVault et al., 2015), *DND* (Lewis et al., 2017), *CA* (Chawla et al., 2021b), and *JI* (Yamaguchi et al., 2021). Table 1 provides an overview of these datasets. Each dataset captures a distinct scenario but is based on the Multi-Issue Bargaining Task (MIBT) framework (Fershtman, 1990), a standard form used for negotiation research, which involves bilateral negotiations over fixed quantities of issues (Table 1). Players have predefined preferences, or *values*, for each issue, which establish their priority order (ex., Scores: 5 = high, 4 = medium, 3 = low priority). In line with standard practice for experimental settings analyzing final outcomes and player behavior, these preferences remain fixed throughout the negotiation.

In MIBT, players engage in dialogue to divide available items to maximize their total points, computed by the inner product of preferences and acquired items in the final deal. MIBT captures various real-world negotiations and has been widely adopted in NLP, beyond NLP (Mell and Gratch, 2017; Acharya and Ortner, 2013; Lau et al., 2008), and industry as well (e.g., iDecisionGames[1]).

Some datasets, like *CA*, report subjective metrics such as *satisfaction from the outcome* and *likeness towards their partner*. Others include utterance-level dialogue act (e.g., *CRA*) and strategy (e.g.,

[1]https://idecisiongames.com/promo-home

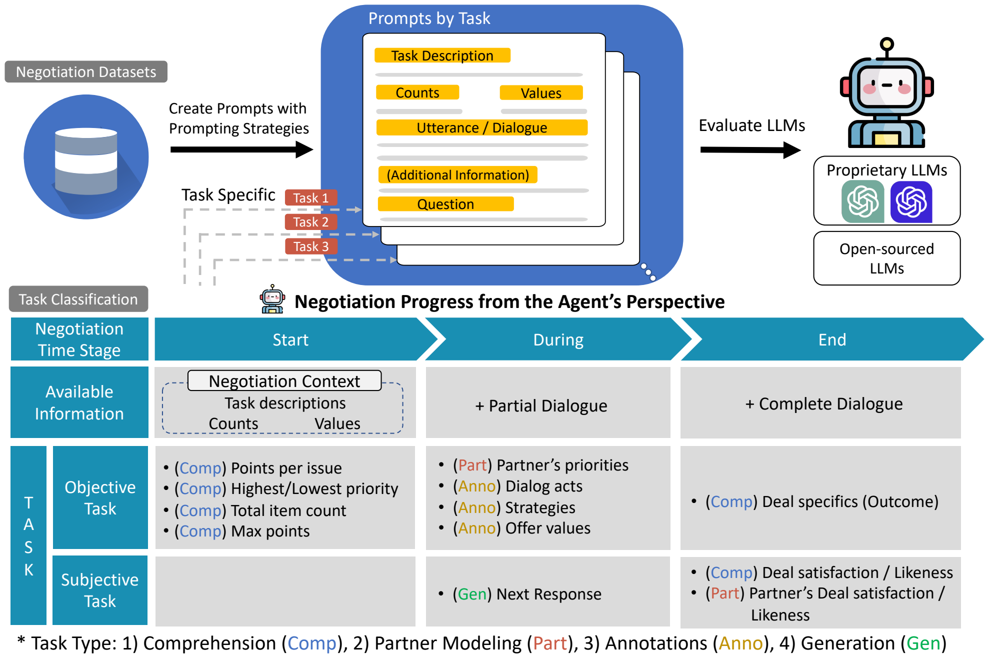


Figure 2: Our methodology for systematically evaluating LLMs in negotiation dialogues. Part A (top) describes the pipeline for creating task-specific prompts from a negotiation dataset and evaluating various LLMs with them. Part B (bottom) depicts the tasks categorized by *Objectivity*, *Time Stage*,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

Dataset	Scenario	Issues	# of Dialogues
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. Depending on the dataset and task, we sample and build test instances per task (capped at 200) at either the utterance level or the dialogue level for our evaluation. The training data is used for fine-tuning our baseline model.

CA) annotations. We use this metadata to design a variety of tasks to evaluate LLMs.

2.2 Task Design

We build 35 tasks based on the aforementioned datasets, guided by the principle that *Every task is designed from the perspective of an agent negotiating for itself*. This is crucial as it governs the information used in the prompts (e.g., explicit partner preferences are not visible).

To categorize these tasks systematically, we use three criteria: 1) *Objectivity*, 2) *Time stage*, and 3) *Task Type* (Figure 2). While tailored for specific negotiation scenarios, these criteria also apply to traditional task-oriented and open-domain dialogues, making our evaluation framework broadly applicable. Detailed descriptions of each task are provided in Table 6 in the Appendix A.

2.2.1 Objectivity

Most tasks are *objective*, testing whether models can accurately answer questions about the objective facts of a negotiation and are evaluated using standardized metrics. In contrast, a few tasks are *subjective* where there is no one correct answer. These include predicting outcome satisfaction and partner likability as reported by human players in post-surveys. Subjective measures are crucial for successful task outcomes, especially in repeated interactions between stakeholders, as they affect user perception of agents (Oliver et al., 1994; Mell et al., 2019; Chawla et al., 2021a). Therefore, we design tasks to assess whether model predictions align with human-reported outcomes by computing correlations. Response generation is also considered as a subjective task, evaluated through automatic and human assessments.

2.2.2 Time Stage

A negotiation dialogue typically unfolds in three stages: *Start*, *During*, and *End*. This distinction impacts the information used in prompts, as an agent can only access information available at each stage from its own perspective.

At the *start*, before any dialogue occurs, the agent knows only the basic information from the negotiation context, including the scenario description, item counts, and priority values (Figure 2). *During* the dialogue, the agent has access to the partial historical conversation. At the *end*, tasks use the entire dialogue within their prompts.

2.2.3 Task Type

Based on prior research, we categorize all subtasks into four types (i.e., Comprehension (Twitchell et al., 2013; Nouri et al., 2013), Annotation (Heddaya et al., 2023), Generation (Lewis et al., 2017), and Partner modeling (Zhang et al., 2020; Chawla et al., 2022)), aligning with traditional higher-level modeling tasks.

Comprehension: These tasks assess the model's ability to understand the negotiation context and outcome, which is fundamental for any sensible negotiation agent (Cao et al., 2015). In the *Start* stage, tasks include identifying the total number of items, issue priorities, and maximum points the agent can gain. In the *End* stage, they evaluate the model's understanding of the final outcomes, including the final deal details (Twitchell et al., 2013) and subjective measures of satisfaction and partner likability (Curhan et al., 2010). Proficiency in these tasks aids in automatically evaluating human-human or human-agent negotiations.

Annotation: These tasks involve annotating utterances with their semantics and functions, such as dialogue acts (e.g., disagree, propose) and linguistic strategies (e.g., elicit-preference, self-need), as well as parsing incoming offers (Chawla et al., 2021b). Annotation is crucial for the Natural Language Understanding (NLU) module in a modular dialogue system, especially *during* the negotiation to understand the partner's utterances and decide the agent's next actions.

Partner Modeling: This is vital *during* the negotiation for understanding the other party's strategy and priorities, and adapting accordingly for favorable outcomes. These tasks are closely related to evaluating the ToM ability of LLMs in inferring the mental states of the partner during negotiations. We focus on predicting the partner's priorities and subjective assessments, such as satisfaction with an outcome and likability towards the other party.

Generation: This task involves generating responses based on context and dialogue history, including crafting offers and counteroffers, responding to incoming offers, and other communicative elements that advance the negotiation. 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 Building Task-specific Prompts

To assess LLMs' capabilities on our proposed tasks, we create task-specific prompts using a standardized template. As shown in Figure 2, the zero-shot prompt template includes five elements: 1) task description, 2) issue counts and values, 3) dialogue or utterance, 4) additional information, and 5) the question. For each task, each element is filled with relevant information from the dataset instance to complete the task-specific prompt. The inclusion of dialogue and utterances depends on the time stage and task definition, while additional information like dialogue acts and strategy types is included only for annotation tasks.

For our analysis in Section 5, we also explore alternative prompting strategies, such as few-shot prompting and Chain-of-Thought (CoT) (Wei et al., 2023). Detailed task descriptions, including questions and applicable datasets, and prompt examples are provided in Appendix A and D, respectively.

3 Experiment Design

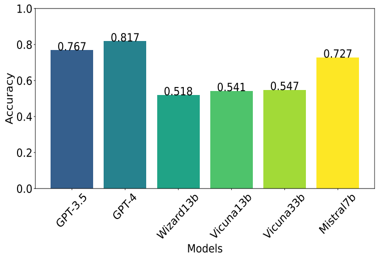
Our primary goal is to analyze the effectiveness of state-of-the-art LLMs for strategic negotiation interactions. Here, we discuss the compared models and evaluation metrics for the results presented in Section 4 (overall zero-shot results) and Section 5 (task-specific analysis to gain further insights).

3.1 Baselines

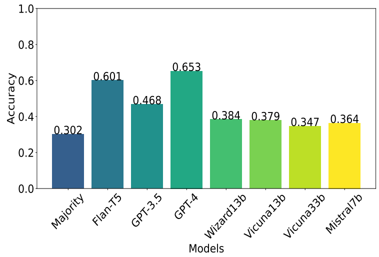
We use the Majority-label voting model and a task-wise fine-tuned Flan-T5 (base) (Chung et al., 2022) as baselines. Flan-T5 was chosen for its strong performance across various NLP tasks and its flexibility in handling diverse input and output formats. For fine-tuning, we use the AdamW optimizer (Loshchilov and Hutter, 2019) with an initial learning rate of 5e-5 and a linear scheduler. Models are trained for up to 5 epochs with a batch size of 8, selecting the best-performing checkpoint for evaluation. Detailed information about the training process, including data processing steps and computational resources, is provided in Appendix B.

3.2 LLMs for evaluation

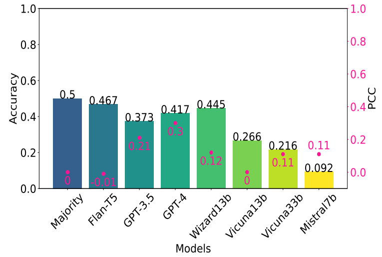
We compare a variety of LLMs, top-performing on popular leaderboards at the time of experimenta-



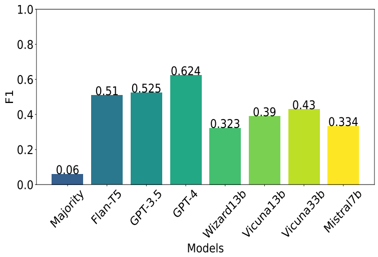
(a) Comprehension (Start) Tasks



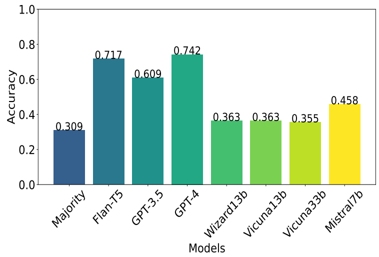
(b) Comprehension (End) Tasks



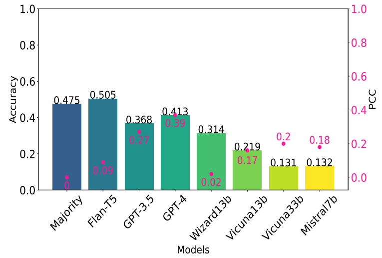
(c) Comprehension (Subjective) Tasks



(d) Annotation Tasks



(e) Partner Modeling Tasks



(f) Partner Modeling (Subjective) Tasks

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. The tasks for these plots have been carefully selected to ensure a fair comparison, with all models passing generation validity checks (i.e., without null values across models), and details of validity check and full results are in Table 9 of Appendix A.

tion. 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 used depend on the task type. For most tasks, primarily Comprehension and Partner modeling tasks, we rely on Accuracy. For annotation tasks, we use Macro-F1 to account for label imbalance. For subjective tasks measuring satisfaction and likeness scores, we use the Pearson Correlation Coefficient (PCC), along with Accuracy. For response generation, we report automatic evaluation with BLEU, ROUGE, and BERTScore[2] (Zhang et al., 2019), and human evaluation by five expert annotators using two subjective metrics on a scale of 1-5: *Coherence* (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 performance?). The metrics used for the tasks are detailed in Table 9 in Appendix A.

[2]https://huggingface.co/google-bert/bert-base-uncased

4 Results

Figure 3 and Table 2 present the key overall trends for zero-shot evaluation on the test set. As detailed in the appendix C, our test set is statistically powered and large enough to ensure reliable comparisons across models. We provide the complete task-wise results in Appendix A and discuss the key findings below.

4.1 Comprehension Tasks

As shown in Figure 3a, GPT-4 outperforms all other models on Comprehension tasks in the *Start* stage, with an average accuracy of 81%. This is followed by GPT-3.5 and Mistral7B, both scoring above 70%. However, these tasks primarily involve questions about the explicitly provided negotiation context, so a simple rule-based parser could achieve 100% accuracy, indicating significant room for improvement.

The *End* stage tasks are more challenging, requiring models to comprehend the dialogue, reason to extract the agreed deal, and compute the answer. As shown in Figure 3b, most models fail to perform well, only marginally better than the trivial Majority baseline. GPT-4 still demonstrates superior performance, surpassing Flan-T5. In comparison, since all evaluated instances end in an *unambiguous* final deal, an expert human can achieve nearly

Model	DND					CA				
	BLEU↑	Rouge-L↑	BERTScore↑	Coherence↑	Strategy↑	BLEU↑	Rouge-L↑	BERTScore↑	Coherence↑	Strategy↑
Human				**4.5**	**4.39**				**4.14**	**3.38**
Flan-T5	**.167**	**.453**	**.678**	4.26*	4.18	**.028**	**.165**	.468	3.21*	2.79*
Mistral7b	.010	.130	.401	3.48*	2.96*	.010	.130	.401	2.99*	2.68*
Wizard13b	.032	.190	.451	3.14*	3.01*	.017	.135	.466	3.08*	2.88*
Vicuna13b	.022	.172	.486	3.48*	3.34*	.015	.135	.472	3.36*	2.92*
Vicuna33b	.038	.216	.547	3.86*	3.74*	.016	.147	.483	3.96	3.06*
GPT-3.5	.030	.200	.467	3.8*	3.50*	.025	.162	**.495**	3.60*	3.01*
GPT-4	.017	.178	.489	4.47	4.04*	.011	.149	.48	4.05	3.24

Table 2: Results on response generation. BLEU, Rouge-L and BERTScore are computed on the full test set (200 instances). Coherence and Strategy are based on a human evaluation of a random subset of 50 examples. * means significantly worse performance than the Human reference according to a Paired Bootstrap test (Sakai, 2006).

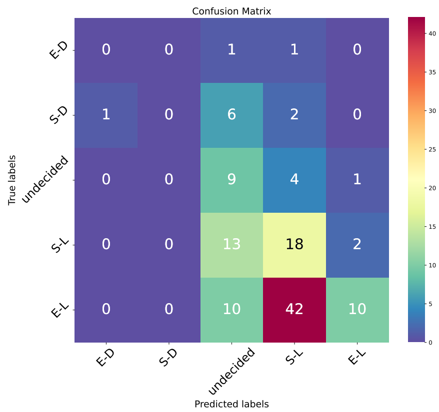


Figure 4: Confusion matrix of predictions of GPT-4 for the subjective task (*end_partner_deal_likeness_ca*). *E* stands for "Extremely", *S* for "Slightly", *D* for "Dislike" and *L* for "Like."

perfect scores on these tasks.

For subjective tasks, we analyze if the LLMs' self-assessments of outcome satisfaction and partner likeness align with those reported by human players. Figure 3c shows that all models perform poorly in terms of Accuracy and PCC. Although GPT-4 exhibited relatively better performance, it sometimes misclassified deal satisfaction in a completely opposite way to humans (e.g., classifying satisfied deals as dissatisfied, or vice versa) (Figure 7 in the Appendix). This inconsistency resulted in a moderate correlation of only 0.3 between GPT-4's satisfaction ratings and human ratings. This leaves uncertainties about the LLMs' abilities to capture the psychological states of the human players in negotiations.

4.2 Annotation Tasks

Figure 3d shows that both GPT-3.5 and GPT-4 outperform the fine-tuned Flan-T5 baseline on annotation tasks, achieving Macro-F1 scores of 62.4% and 52.5%, respectively. Task-wise results (Table 9 in Appendix A) indicate that these trends are influenced by Flan-T5's poor performance on the *dur_dial_act_ji* task. However, Flan-T5 performs better than GPT-4 on most tasks, highlighting room for improvement in zero-shot scenarios. Additionally, models struggle more with detecting negotiation strategies than dialogue acts, likely due to the subtlety of strategy expressions. These results are without few-shot examples or prior utterances, which we will explore further in Section 5.

4.3 Partner Modeling Tasks

Figure 3e presents the results for objective partner modeling tasks (inferring partner priorities) with the entire dialogue included in the prompt. GPT-4 achieves the best performance, demonstrating strong out-of-the-box Theory of Mind (ToM) abilities, unlike smaller open-source variants that mostly perform similarly to the Majority baseline. In Section 5, we further analyze how model performance varies with the number of utterances seen by the model.

For inferring the partner's subjective assessments (Figure 3f), Flan-T5 achieves an accuracy of 50.5%, while GPT-4 scores the highest PCC at 0.39, highlighting the generally poor performance of models in assessing subjective perceptions. The skewed distribution in the confusion matrix (Figure 4) shows GPT-4 frequently predicts neutral or slight favorability in over 50% of cases where partners demonstrate strong likeness, suggesting that GPT-4 fails to capture the degree of positivity that humans display. Prior work suggests that including partner demographics, personality, and emotional expression in the dialogue can improve these predictions (Chawla et al., 2023a).

4.4 Generation Tasks

Table 2 shows that Flan-T5 significantly outperforms other models on BLEU and ROUGE, likely due to dataset-specific fine-tuning. LLMs strug-

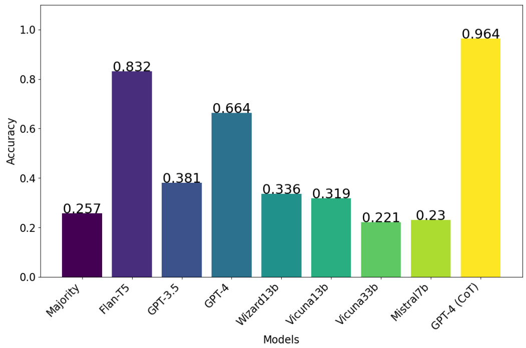
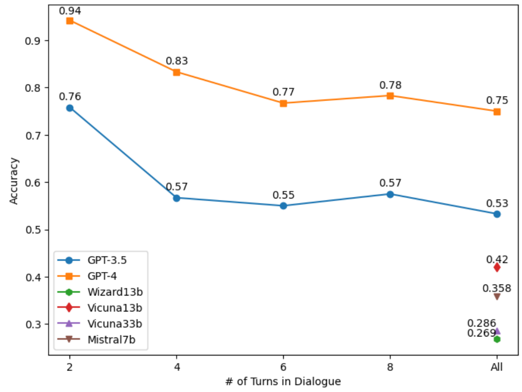
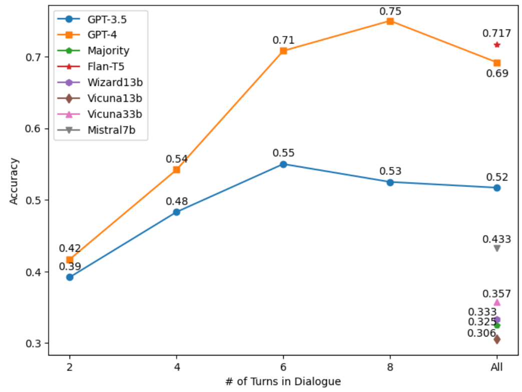


Figure 5: GPT-4's evaluation on the *end_deal_total_dnd* task, highlighting the impact of Chain-of-Thought (CoT) prompting. Results for other tasks can be found in Figure 9 in the appendix.



(a) *dur_ask_low_priority_ca*



(b) *dur_partner_ask_low_priority_ca*

Figure 6: Evaluation on the tasks demonstrating the impact of increasing the number of utterances seen by the models.

gle to align with the dataset's utterance style and structure, leading to low overlap with reference (examples in Appendix F). However, on the CA dataset, which is richer and more diverse than DND, GPT models slightly surpass fine-tuned T5 in BERTScore, demonstrating better semantic understanding and generating responses semantically more similar to human reference.

Automatic evaluation has limitations in accurately assessing the appropriateness of model responses in complex negotiation scenarios. Therefore, we additionally conducted an expert human evaluation on a subset of 50 examples and confirmed a high inter-rater agreement with an Intra-Class Correlation Coefficient (ICC) (Koo and Li, 2016) of 0.82. The human reference baseline performs the best on both *Coherence* and *Strategy* for both datasets. However, GPT-4 showed comparable *Coherence* scores to humans, indicating it performs similarly to humans. In *Strategy*, only the GPT-4 model approached human performance on the *CA* dataset, but in all other cases, all models significantly underperformed compared to human references. Further, the models perform better on *Coherence* 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.

Although GPT-4's performance is comparable to humans in our tasks, it is *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., 2023c) relies on a combination of supervised learning (learning average behavior) and RL (explicit training for task-specific rewards). Therefore, we perform an error analysis and discuss key limitations to guide future work.

Sophisticated *Incoherent* Generations: Handling off-topic interaction remains challenging, and trivial issues like repetition still persist. We also found sophisticated coherence errors. For example, given the dialogue history from *DND*, "*...YOU: I can't accept that. how about I get the book and 1 hat, and you can get the balls and 1 hat? THEM: I can't accept that deal... can I have one of each ?*", 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,*

Tasks	Label-Balance (Tasks Difficulty)	Metric	Model					
			GPT-3.5	GPT-4	Mistral7b	Vicuna13b	Vicuna33b	Wizard13b
High / Low Priority Tasks	Well-Balacned (Easy)	Acc. ↑	0.677	**0.91**	0.419	0.206	0.243	0.38
		F1 ↑	0.669	**0.908**	0.328	0.14	0.115	0.348
		|Acc. - F1| ↓	0.007	**0.002**	0.091	0.066	0.128	0.032
		KL-D↓	0.189	**0.02**	0.626	1.134	1.286	0.372
	Well-Balacned (Hard)	Acc.↑	0.638	**0.825**	0.458	0.538	0.331	0.307
		F1↑	0.623	**0.824**	0.453	0.53	0.215	0.184
		|Acc. - F1| ↓	0.015	**0.001**	0.006	0.008	0.116	0.123
		KL-D↓	0.08	**0.019**	0.069	0.046	0.978	1.038
Dial-Act / Strategy Tasks	Imbalanced (Hard)	Acc↑	0.853	**0.898**	0.742	0.71	0.81	0.641
		F1 ↑	0.525	**0.624**	0.334	0.39	0.43	0.323
		|Acc - F1| ↓	0.328	**0.274**	0.408	0.32	0.38	0.318
		KL-D↓	**0.732**	0.951	1.952	1.442	1.13	1.552

Table 3: Evaluation results of LLMs on balanced and imbalanced tasks. The KL-D metric represents KL-Divergence. The reported values are average metrics of the tasks, selected based on the task type, label type (nominal class), label distribution, and generation validity (i.e., no null values), as described in the table 9 in the Appendix. Further details on the selected tasks can be found in Appendix E.

we both get something we want.". Although the output seems novel, it paraphrases YOU's previous offer, which the partner already rejected. Ensuring coherence remains an open challenge 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, and also evaluate prediction balance and robustness in LLMs.

CoT Prompting: We focus on tasks requiring multi-hop arithmetic reasoning, such as calculating the maximum possible points and inferring actual total points in a negotiation. Using CoT with GPT-4 yields nearly 100% accuracy on the tasks we tested (Figure 5), highlighting its effectiveness for LLMs in negotiation tasks. A sample CoT prompt is provided in Table 13 in Appendix.

Using Prior Utterances for Annotation Tasks: We include two prior utterances as adding additional contexts in prompts for annotation tasks. The results are mixed (Figure 8 in Appendix E)—performance improves in only one task and shows minor degradation in the other three. The type of annotation labels and data collection methodology likely influence these outcomes, and irrelevant utterances might confuse the models. Including prior utterances should be a domain-specific choice based on validation performance.

Few-shot Prompting: In-Context Learning (ICL) is a key ability of LLMs, enabling them to better understand tasks and follow instructions (Wei et al., 2022). We examine whether adding two randomly sampled examples to annotation tasks enhances performance via ICL. Our findings show that ICL improves performance on two out of four tasks, indicating that few-shot examples can boost model performance (Figure 8 in Appendix E). Selecting optimal examples for ICL remains an active research area, which could benefit tasks involving strategic interactions.

Varying the number of seen utterances We explore two questions: 1) Do additional utterances confuse the model about its own context provided in the prompt? and 2) Does the model utilize additional utterances to infer the partner's context?

The answer to both questions is 'Yes' (Figure 6). When asked about its own preferences, model performance degrades with more utterances. However, performance improves when inferring partner preferences with additional utterances. This can be explained by *recency bias*—models focus more on recent information, improving partner modeling but diminishing comprehension of their own context given at the start.

5.1 Evaluating Prediction Balance and Robustness in LLMs

We test LLMs' ability to make balanced predictions to assess model robustness. As shown in Table 3, GPT-4 performs best in both easy and hard well-balanced priority tasks, with the smallest differences in accuracy and F1 score, and near-zero KL-Divergence (KL-D), indicating highly balanced predictions. In imbalanced tasks, although all models show decreased F1 scores relative to accuracy, GPT-4 has the smallest decline and low KL-D, indicating relatively balanced predictions compared to other models. Our results confirm GPT-4's superior robustness compared to other LLMs, as its balanced predictions on unseen data demonstrate enhanced generalization ability.

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. (2023c) 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) argues that their understanding is superficial, and Ullman (2023) provides 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 understanding, inference about the other party's needs, and reasoning for strategic decision-making.

7 Conclusion

We devise a methodology to systematically analyze the multifaceted capabilities of LLMs in negotiations. When evaluated out-of-the-box, GPT-4 outperforms and is more robust than other LLMs but still leaves room for improvement in most tasks. CoT and few-shot prompting help improve performance in several arithmetic reasoning and annotation tasks. In contrast, smaller open-source models struggle, 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, this is primarily true for proprietary LLMs, emphasizing the need to improve smaller open-source models.

We also find that information in dialogue can confuse models about their own context, even when provided in the prompts. Future work should focus on helping LLMs handle longer contexts by emphasizing relevant input. Human evaluation shows all models struggle with generating strategically appropriate responses. Recent efforts in structured reasoning (Zhou et al., 2024) may address this by breaking response generation into subgoals like comprehension, annotation, and partner modeling. We plan to combine LLMs with RL policies from prior negotiation work (He et al., 2018) to control the dialogue agent's strategy or personality.

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. The evaluation of negotiation dialogues requires domain knowledge and sophisticated assessment, hence it was conducted by five expert annotators, including the authors of this work who possess expertise in this field. General crowdsourcing was not ideal for this evaluation.

8.4 AI for Social Influence Interactions

Negotiation dialogues fall under the broader spectrum of social influence tasks (Chawla et al., 2023b), 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)[3] 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 (i.e., dynamic 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.

[3] https://sites.google.com/view/sicon-2023/home

References

Sahar Abdelnabi, Amr Gomaa, Sarath Sivaprasad, Lea Schönherr, and Mario Fritz. 2023. Llm-deliberation: Evaluating llms with interactive multi-agent negotiation games.

Avidit Acharya and Juan Ortner. 2013. Delays and partial agreements in multi-issue bargaining. *Journal of Economic Theory*, 148(5):2150–2163.

Steffen Andersen, Seda Ertac, Uri Gneezy, John A List, and Sandra Maximiano. 2018. On the cultural basis of gender differences in negotiation. *Experimental Economics*, 21:757–778.

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.

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.

Mukun Cao, Xudong Luo, Xin Robert Luo, and Xiaopei Dai. 2015. Automated negotiation for e-commerce decision making: a goal deliberated agent architecture for multi-strategy selection. *Decision Support Systems*, 73:1–14.

Kushal Chawla, Rene Clever, Jaysa Ramirez, Gale Lucas, and Jonathan Gratch. 2021a. Towards emotion-aware agents for negotiation dialogues. In *2021 9th International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 1–8. IEEE.

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. *IEEE Transactions on Affective Computing*.

Kushal Chawla, Gale M Lucas, Jonathan May, and Jonathan Gratch. 2022. Opponent modeling in negotiation dialogues by related data adaptation. *arXiv preprint arXiv:2205.00344*.

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. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3167–3185, Online. Association for Computational Linguistics.

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. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 750–766, Dubrovnik, Croatia. Association for Computational Linguistics.

Kushal Chawla, Ian Wu, Yu Rong, Gale Lucas, and Jonathan Gratch. 2023c. Be selfish, but wisely: Investigating the impact of agent personality in mixed-motive human-agent interactions. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13078–13092.

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.

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.

Jared R Curhan, Hillary Anger Elfenbein, and Noah Eisenkraft. 2010. The objective value of subjective value: A multi-round negotiation study. *Journal of Applied Social Psychology*, 40(3):690–709.

Janez Demšar. 2006. Statistical comparisons of classifiers over multiple data sets. *The Journal of Machine learning research*, 7:1–30.

David DeVault, Johnathan Mell, and Jonathan Gratch. 2015. Toward natural turn-taking in a virtual human negotiation agent. In *2015 AAAI Spring Symposium Series*.

Chaim Fershtman. 1990. The importance of the agenda in bargaining. *Games and Economic Behavior*, 2(3):224–238.

Chris Frith and Uta Frith. 2005. Theory of mind. *Current biology*, 15(17):R644–R645.

Yao Fu, Hao Peng, Tushar Khot, and Mirella Lapata. 2023. Improving language model negotiation with self-play and in-context learning from ai feedback.

Kanishk Gandhi, Dorsa Sadigh, and Noah D. Goodman. 2023. Strategic reasoning with language models.

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.

He He, Derek Chen, Anusha Balakrishnan, and Percy Liang. 2018. Decoupling strategy and generation in negotiation dialogues. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2333–2343.

Mourad Heddaya, Solomon Dworkin, Chenhao Tan, Rob Voigt, and Alexander Zentefis. 2023. Language of bargaining.

Jie Huang and Kevin Chen-Chuan Chang. 2023. Towards reasoning in large language models: A survey.

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.

Emmanuel Johnson, Gale Lucas, Peter Kim, and Jonathan Gratch. 2019. Intelligent tutoring system for negotiation skills training. In *International Conference on Artificial Intelligence in Education*, pages 122–127. Springer.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2023. Large language models are zero-shot reasoners.

Terry K Koo and Mae Y Li. 2016. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of chiropractic medicine*, 15(2):155–163.

Michal Kosinski. 2023. Theory of mind might have spontaneously emerged in large language models.

Raymond YK Lau, On Wong, Yuefeng Li, and Louis CK Ma. 2008. Mining trading partners' preferences for efficient multi-issue bargaining in e-business. *Journal of Management Information Systems*, 25(1):79–104.

Roy J Lewicki, Bruce Barry, and David M Saunders. 2016. *Essentials of negotiation*. McGraw-Hill New York.

Mike Lewis, Denis Yarats, Yann N. Dauphin, Devi Parikh, and Dhruv Batra. 2017. Deal or no deal? end-to-end learning for negotiation dialogues.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization.

Peng Luo. 2008. Analysis of cultural differences between west and east in international business negotiation. *International Journal of Business and Management*, 3(11):103–106.

Johnathan Mell and Jonathan Gratch. 2017. Grumpy & pinocchio: answering human-agent negotiation questions through realistic agent design. In *Proceedings of the 16th conference on autonomous agents and multiagent systems*, pages 401–409.

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 *2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 1–7. IEEE.

David Noever and Forrest McKee. 2023. Numeracy from literacy: Data science as an emergent skill from large language models.

Elnaz Nouri, Sunghyun Park, Stefan Scherer, Jonathan Gratch, Peter J Carnevale, Louis-Philippe Morency, and David R Traum. 2013. Prediction of strategy and outcome as negotiation unfolds by using basic verbal and behavioral features. In *INTERSPEECH*, pages 1458–1461.

Richard L Oliver, PV Sundar Balakrishnan, and Bruce Barry. 1994. Outcome satisfaction in negotiation: A test of expectancy disconfirmation. *Organizational Behavior and Human Decision Processes*, 60(2):252–275.

OpenAI. 2022. Openai: Introducing chatgpt.

OpenAI. 2023. Openai: Gpt-4.

Tetsuya Sakai. 2006. Evaluating evaluation metrics based on the bootstrap. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 525–532.

Douglas P Twitchell, Matthew L Jensen, Douglas C Derrick, Judee K Burgoon, and Jay F Nunamaker. 2013. Negotiation outcome classification using language features. *Group Decision and Negotiation*, 22:135–151.

Tomer Ullman. 2023. Large language models fail on trivial alterations to theory-of-mind tasks.

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.

Jason Wei, Xuezhi 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.

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. *arXiv preprint arXiv:2304.12244*.

Atsuki Yamaguchi, Kosui Iwasa, and Katsuhide Fujita. 2021. Dialogue act-based breakdown detection in negotiation dialogues. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 745–757, Online. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Zheng Zhang, Lizi Liao, Xiaoyan Zhu, Tat-Seng Chua, Zitao Liu, Yan Huang, and Minlie Huang. 2020. Learning goal-oriented dialogue policy with opposite agent awareness. *arXiv preprint arXiv:2004.09731*.

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. *arXiv preprint arXiv:2402.03620*.

Caleb Ziems, William Held, Omar Shaikh, Jiaao Chen, Zhehao Zhang, and Diyi Yang. 2023. Can large language models transform computational social science? *arXiv preprint arXiv:2305.03514*.

A Details of Negotiation Tasks

We present an overview and detailed task descriptions of all the designed tasks in Table 5 and Table 6. Table 7 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 8 presents the final list of selected tasks used for computing these plots. Finally, we report the task-wise results for all models in Table 9 for completeness.

B 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, excluding 200 test cases for LLM evaluation, split the remaining data into training and development sets in a 9:1 ratio for model training and evaluation.
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 a 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 model[4] 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 BLEU[5] and ROUGE-L[6] respectively.

C Statistical Power Analysis for Test set

We used a capped test set of 200 samples across tasks. To ensure this sample size was sufficient to detect model differences, we conducted a Chi-squared power analysis. With a significance level of p = 0.05, a medium effect size (W = 0.3), and four outcome categories, the power for 200 samples was 0.959, confirming adequacy for detecting meaningful differences.

We applied the McNemar test to assess model performance (Demšar, 2006), focusing on T5 and GPT-4 across all tasks with valid results (excluding the subjective tasks with ordinal variables) in Table 9. As shown in Table 4 below, when performance was similar, no significant differences were found, but statistically significant differences were detected where the models' outcomes diverged. This confirmed that the sample size was sufficient to detect differences between models.

Task name	Models		McNeMar's test	
	T5	GPT-4	Chi-square (DF:1)	P-value
dur_dial_act_cra	0.787	0.678	15.63	0.0001*
dur_dial_act_dnd	0.96	0.825	5.14	0.0233*
dur_dial_act_ji	0.019	0.578	145.31	0.0001*
dur_full_proposal_cra	0.439	0.369	2.95	0.0859
dur_full_proposal_dnd	1	0.866	79.01	0.0001*
dur_partner_ask_high_priority_ca	0.717	0.792	1.49	0.2225
dur_partner_ask_low_priority_ca	0.717	0.75	0.1	0.7488
dur_strategy_ca	0.724	0.507	4.97	0.0259*
end_deal_specifics_ca	0.364	0.664	177.03	0.0001*
end_deal_specifics_dnd	0.973	0.67	40.45	0.0001*
end_deal_specifics_ji	0.764	0.858	73.29	0.0001*
end_deal_total_ca	0.233	0.083	9.63	0.0019*
end_deal_total_dnd	0.832	0.664	9.26	0.0023*

Table 4: McNeMar's test results for the two models. * indicates statistical significance (P-value < 0.05)

[4] https://huggingface.co/google/flan-t5-base
[5] https://pypi.org/project/evaluate/
[6] https://pypi.org/project/rouge-score/

D 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 10, 11, and 12 provide examples of prompts for specific tasks corresponding to the *start*, *during*, and *end* time stages respectively. Additionally, Table 13 showcases an example of CoT prompting used for the *end_deal_total_ca* task. Table 14 shows the questions used in prompts for all tasks.

E Task-specific Analysis

Figure 7 displays the prediction results of GPT-4 on selected subjective tasks with a label-wise confusion matrix. Figure 8 presents the results for few-shot prompting and incorporating previous utterances for various *Annotation* tasks. Figure 9 shows the results for the Chain-of-Thought (CoT) prompting technique on four multi-hop arithmetic reasoning *Comprehension* tasks. We also present the effect of varying the number of utterances seen by the models for preference prediction tasks in Figure 10.

Table 15 shows the selected tasks in Table 3 presenting the evaluation results of LLMs on both well-balanced and imbalanced tasks.

Task Type	**Tasks**
Hard Priority Tasks	*mid_partner_ask_high/low_priority_ca*
Easy Priority Tasks	*mid_ask_high/low_priority_ji, sta_ask_high/low_priority_ji*
Dialog-Act/Strategy Tasks	*mid_strategy_ca, mid_dial_act_cra, mid_dial_act_dnd*

Table 15: Selected tasks categorized into Hard and Easy Priority Tasks, and Dialog-Act/Strategy Tasks.

F Generation Outputs

We present sample generations by the compared models along with the ground-truth reference in Tables 16 and 17.

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 5: 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., 2021b), *JI*: (Yamaguchi et al., 2021).

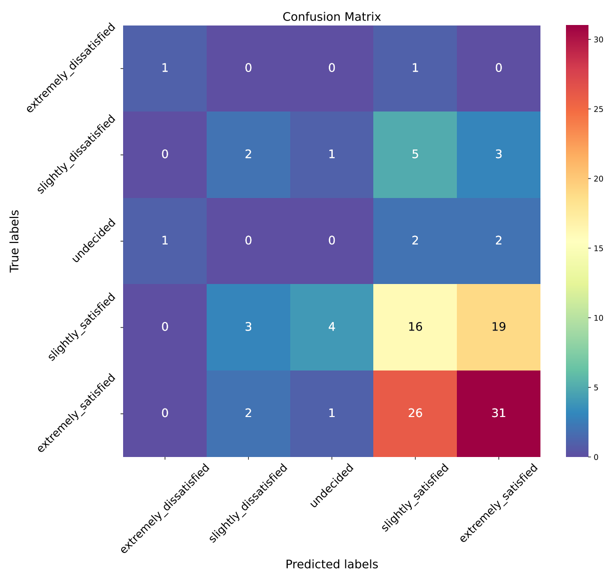


Figure 7: Confusion matrix of predictions of GPT-4 for the subjective task (*end_deal_satisfaction_ca*).

Task Name	Task Descripttion
sta_total_item_count_dnd	In the Start Stage of negotiation in the DND dataset, the task involves the Agent accurately understanding the count of items that can be acquired in a negotiation, given the negotiation conditions.
sta_total_item_count_ca	In the Start Stage of negotiation in the CA dataset, the task involves the Agent accurately understanding the count of items that can be acquired in a negotiation, given the negotiation conditions.
sta_max_points_dnd	In the Start Stage of negotiation in the DND dataset, the task involves the Agent accurately understanding the maximum score that can be achieved in a negotiation, given the negotiation conditions.
sta_max_points_ca	In the Start Stage of negotiation in the CA dataset, the task involves the Agent accurately understanding the maximum score that can be achieved in a negotiation, given the negotiation conditions.
sta_ask_point_values_dnd	In the Start Stage of negotiation in the DND dataset, the task involves the Agent accurately understanding its own value (i.e., priority) of each item, given the negotiation conditions.
sta_ask_point_values_ca	In the Start Stage of negotiation in the CA dataset, the task involves the Agent accurately understanding its own value (i.e., priority) of each item, given the negotiation conditions.
sta_ask_low_priority_ji	In the Start Stage of negotiation in the JI dataset, the task involves the Agent accurately understanding its least prioritized item, given the negotiation conditions.
sta_ask_low_priority_ca	In the Start Stage of negotiation in the CA dataset, the task involves the Agent accurately understanding its least prioritized item, given the negotiation conditions.
sta_ask_high_priority_ji	In the Start Stage of negotiation in the JI dataset, the task involves the Agent accurately understanding its most prioritized item, given the negotiation conditions.
sta_ask_high_priority_ca	In the Start Stage of negotiation in the CA dataset, the task involves the Agent accurately understanding its most prioritized item, given the negotiation conditions.
dur_strategy_ca	In the During Stage of negotiation in the CA dataset, the task involves annotating negotiation strategies for a specific utterance in a negotiation dialogue.
dur_partner_ask_low_priority_ji	In the During Stage of negotiation in the JI dataset, the task involves the Agent inferring the partner's least prioritized item from the given negotiation dialogue.
dur_partner_ask_low_priority_ca	In the During Stage of negotiation in the CA dataset, the task involves the Agent inferring the partner's least prioritized item from the given negotiation dialogue.
dur_partner_ask_high_priority_ji	In the During Stage of negotiation in the JI dataset, the task involves the Agent inferring the partner's most prioritized item from the given negotiation dialogue.
dur_partner_ask_high_priority_ca	In the During Stage of negotiation in the CA dataset, the task involves the Agent inferring the partner's most prioritized item from the given negotiation dialogue.
dur_gen_resp_dnd	In the During Stage of negotiation in the DND dataset, the task involves the Agent generating an appropriate next response from the given negotiation dialogue.
dur_gen_resp_ca	In the During Stage of negotiation in the CA dataset, the task involves the Agent generating an appropriate next response from the given negotiation dialogue.
dur_full_proposal_dnd	In the During Stage of negotiation in the DND dataset, the task involves annotating a full offer (i.e., counts of each item in the offer) from a specific utterance in a negotiation dialogue.
dur_full_proposal_cra	In the During Stage of negotiation in the CRA dataset, the task involves annotating a full offer (i.e., count of each item in the offer) from a specific utterance in a negotiation dialogue.
dur_dial_act_ji	In the During Stage of negotiation in the JI dataset, the task involves annotating dialogue acts for a specific utterance in a negotiation dialogue.
dur_dial_act_dnd	In the During Stage of negotiation in the DND dataset, the task involves annotating dialogue acts for a specific utterance in a negotiation dialogue.
dur_dial_act_cra	In the During Stage of negotiation in the CRA dataset, the task involves annotating dialogue acts for a specific utterance in a negotiation dialogue.
dur_ask_low_priority_ji	In the During Stage of negotiation in the JI dataset, the task involves the Agent accurately understanding its least prioritized item from the given negotiation dialogue.
dur_ask_low_priority_ca	In the During Stage of negotiation in the CA dataset, the task involves the Agent accurately understanding its least prioritized item from the given negotiation dialogue.
dur_ask_high_priority_ji	In the During Stage of negotiation in the JI dataset, the task involves the Agent accurately understanding its most prioritized item from the given negotiation dialogue.
dur_ask_high_priority_ca	In the During Stage of negotiation in the CA dataset, the task involves the Agent accurately understanding its most prioritized item from the given negotiation dialogue.
end_partner_deal_satisfaction_ca	In the End Stage of negotiation in the CA dataset, the task involves the Agent inferring the final deal satisfaction of the partner from the given negotiation dialogue.
end_partner_deal_likeness_ca	In the End Stage of negotiation in the CA dataset, the task involves the Agent inferring the partner's likeness towards itself from the given negotiation dialogue.
end_deal_total_dnd	In the End Stage of negotiation in the DND dataset, the task involves the Agent understanding the final score of the deal (i.e., the inner product of item counts and values) from the given negotiation dialogue.
end_deal_total_ca	In the End Stage of negotiation in the CA dataset, the task involves the Agent understanding the final score of the deal (i.e., the inner product of item counts and values) from the given negotiation dialogue.
end_deal_specifics_ji	In the End Stage of negotiation in the JI dataset, the task involves the Agent understanding the details of the final deal (i.e., item counts of each item in the deal) from the given negotiation dialogue.
end_deal_specifics_dnd	In the End Stage of negotiation in the DND dataset, the task involves the Agent understanding the details of the final deal (i.e., item counts of each item in the deal) from the given negotiation dialogue.
end_deal_specifics_ca	In the End Stage of negotiation in the CA dataset, the task involves the Agent understanding the details of the final deal (i.e., item counts of each item in the deal) from the given negotiation dialogue.
end_deal_satisfaction_ca	In the End Stage of negotiation in the CA dataset, the task involves the Agent understanding its own final deal satisfaction from the given negotiation dialogue.
end_deal_likeness_ca	In the End Stage of negotiation in the CA dataset, the task involves the Agent understanding its likeness towards the partner from the given negotiation dialogue.

Table 6: Task descriptions of all designed tasks for evaluating LLMs

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 7: 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 8: 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	Acc./PCC	0.525/0	0.525/0	0.357/0.419	0.175/0.367		0.119/-0.033	0.267/0.245	0.239/0.234
end_deal_satisfaction_ca	Acc./PCC	0.5/0	0.467/-0.008	0.373/0.211	0.417/0.304	0.092/0.111	0.266/0.001	0.216/0.114	0.445/0.118
end_deal_specifics_ca	Acc.	0.356	0.364	0.664	0.916	0.517	0.517	0.593	0.555
end_deal_total_ca	Acc.	0.142	0.233	0.158	0.083	0.15	0.05	0.017	0.017
end_partner_deal_likeness_ca	Acc./PCC	0.517/0	0.517/0	0.31/0.295	0.308/0.423	0.133/0.102	0.167/0.259	0.178/0.283	0.282/-0.086
end_partner_deal_satisfaction_ca	Acc./PCC	0.433/0	0.492/0.181	0.426/0.236	0.517/0.36	0.13/0.26	0.271/0.08	0.083/0.114	0.345/0.124
dur_ask_high_priority_ca	Acc.			0.742	0.9	0.558		0.375	0.345
dur_ask_low_priority_ca	Acc.			0.533	0.75	0.358		0.286	0.269
dur_partner_ask_high_priority_ca	Acc.	0.292	0.717	0.7	0.792	0.483	0.42	0.353	0.392
dur_partner_ask_low_priority_ca	Acc.	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	Acc.			1	1				0.667
sta_ask_low_priority_ca	Acc.			1	1	0.5			0.4
sta_ask_point_values_ca	Acc.			1	1	1	1	1	1
sta_max_points_ca	Acc.			0.333	0.333	0.5	0	0	0
sta_total_item_count_ca	Acc.			1	1	1	1	1	0.333
dur_dial_act_cra	F1	0.067	0.787	0.535	.678	0.35	0.338	0.518	0.302
dur_full_proposal_cra	Acc.	0.359	0.439	0.352	0.369	0.241	0.262	0.245	0.325
end_deal_specifics_dnd	Acc.	0.454	0.973	0.67	0.949	0.558	0.631	0.558	0.628
end_deal_total_dnd	Acc.	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	Acc.	0.39	1	0.742	0.866	0.648	0.748	0.725	0.687
sta_ask_point_values_dnd	Acc.			0.993	1	1	1	0.752	1
sta_max_points_dnd	Acc.			0.317	0.337	0.366	0.495	0.307	0.386
sta_total_item_count_dnd	Acc.			0.95	1	0.98	0.505	0.901	0.465
end_deal_specifics_ji	Acc.	0.261	0.764	0.782	0.858	0.733	0.8	0.785	0.766
dur_ask_high_priority_ji	Acc.			0.495	0.862	0.37	0.233	0.252	0.259
dur_ask_low_priority_ji	Acc.			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	Acc.	0.165	0.202	0.193			0.198	0.204	0.204
dur_partner_ask_low_priority_ji	Acc.	0.193	0.266	0.202		0.269	0.176	0.157	0.13
sta_ask_high_priority_ji	Acc.			0.78	0.89	0.505	0.155	0.211	0.596
sta_ask_low_priority_ji	Acc.			0.761	0.972	0.468	0.174	0.202	0.367

Table 9: 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 for other LLMs indicate that the model failed to produce the intended or valid output for the given task-specific prompt, not passing our generation validity check (threshold: 80% valid response rate), 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 10: Prompt example of the Start-stage Task: *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 11: Prompt example of the During-stage Task: *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: Hmmm...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 12: Prompt example of the End-stage Task: 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>

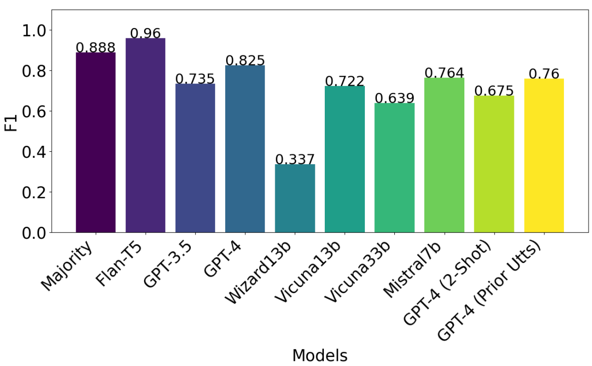
Here are the number of points you get for each type of package, contained in <value> tags.
<value>
Each Food Package: 3 points
Each Water Package: 5 points
Each Firewood Package: 4 points
</value>

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: Hmmm...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>

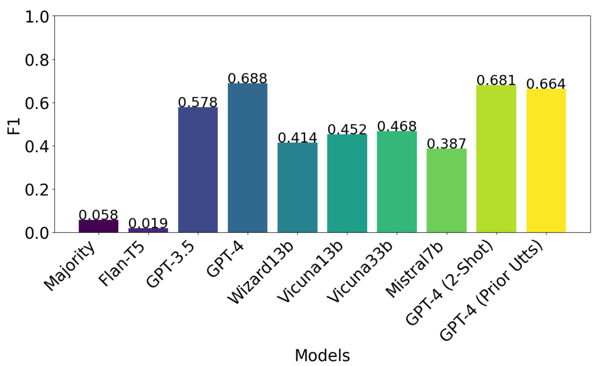
Question: How many points did you get at the end of the negotiation?

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.

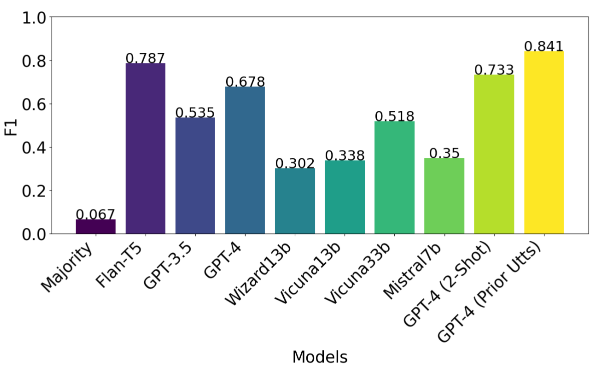
Table 13: CoT prompt example for the *end_deal_total_ca* task.



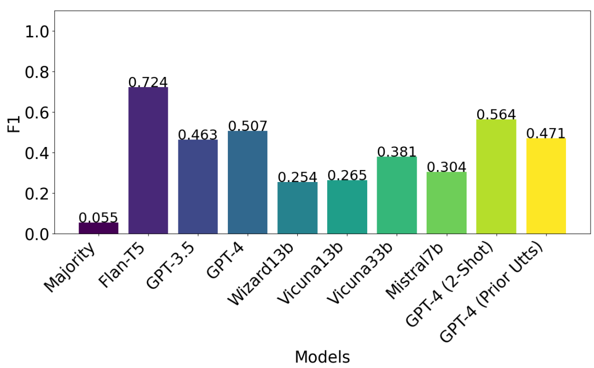
(a) Evaluation on *dur_dial_act_dnd* task



(b) Evaluation on *dur_dial_act_ji* task



(c) Evaluation on *dur_dial_act_cra* task



(d) Evaluation on *dur_strategy_ca* task

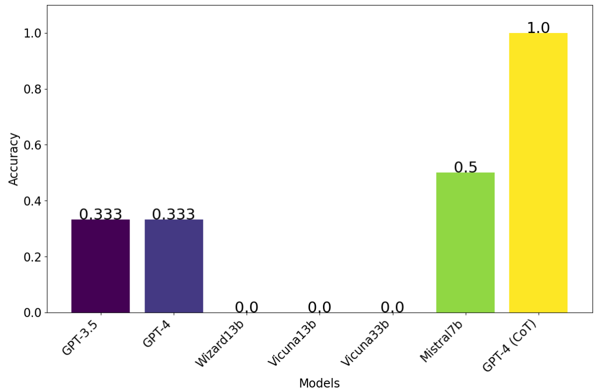
Figure 8: Evaluation on four tasks demonstrating the impact of 1) Two-shot prompting and 2) Incorporating two prior utterances. We only consider GPT-4 for this analysis. F1 refers to the macro F1 score over all labels.

Task	Question
sta_total_item_count_dnd *sta_total_item_count_ca*	What is the total number of items being negotiated over? Present your answer as a single number with no additional text.
sta_max_points_dnd *sta_max_points_ca*	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.
sta_ask_point_values_dnd	How many points is one item of each issue worth to you? Present your answer as a JSON within <answer></answer>tags with keys as issues (books, hats, and balls) and values as the corresponding answers.
sta_ask_point_values_ca	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.
sta_ask_low_priority_ji *dur_ask_low_priority_ji*	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
sta_ask_low_priority_ca *dur_ask_low_priority_ca*	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
sta_ask_high_priority_ji *dur_ask_low_priority_ca*	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
sta_ask_high_priority_ca *dur_ask_high_priority_ca*	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
dur_strategy_ca	Which negotiation strategies are employed in the utterance? Present your answer as a comma-separated list of strategies, contained in <answer></answer>tags with no additional text.
dur_partner_ask_low_priority_ji	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
dur_partner_ask_low_priority_ca	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
dur_partner_ask_high_priority_ji	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
dur_partner_ask_high_priority_ca	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
dur_gen_resp_dnd *dur_gen_resp_ca*	Given the recent dialogue history inside <dialogue>tags, generate your next response in the negotiation concisely, following a similar style as previous utterances.
dur_full_proposal_dnd	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, pick your best guess.
dur_full_proposal_cra	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 (painting, lamp, and record) and values as the corresponding answers. If the answer is not clear for an issue, output NA.
dur_dial_act_ji *dur_dial_act_cra*	Which dialogue acts are employed in the utterance delimited by the<utterance>tags? Present your answer as a Python list of the relevant options. At least one option applies.
dur_dial_act_dnd	Which dialogue act is employed in the utterance contained in<utterance>tags? Present your answer as a single word.
end_partner_deal_satisfaction_ca	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
end_partner_deal_likeness_ca	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
end_deal_total_dnd *end_deal_total_ca*	How many points did you get at the end of the negotiation? Present your answer as a single number with no additional text.
end_deal_specifics_ji	In the final deal, what value was agreed on for each issue? Present your answer as a JSON within <answer></answer>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.
end_deal_specifics_dnd	In the final deal, how many items of each issue did you get? 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 you are unsure, pick your best guess.
end_deal_specifics_ca	In the final deal, how many items 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 you are unsure, pick your best guess.
end_deal_satisfaction_ca	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
end_deal_likeness_ca	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

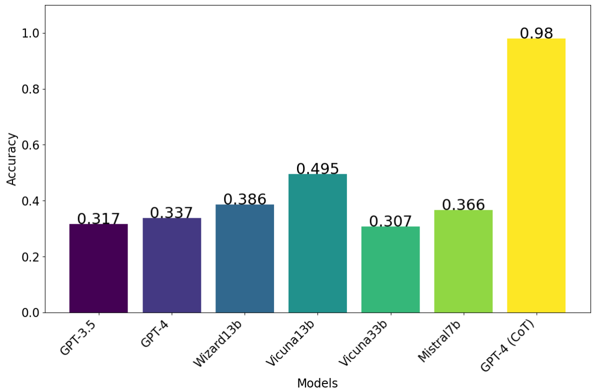
Table 14: Questions used in the prompts for all tasks.

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	Coherence	Strategy
Reference	YOU: deal !	5	5
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

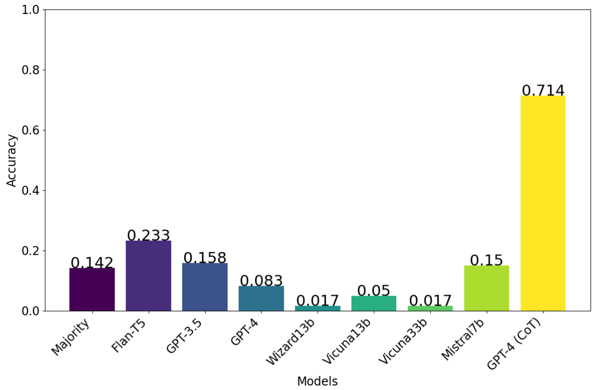
Table 16: Sample responses generated by the models for an instance from the DND dataset. The *Coherence* and *Strategy* scores are based on human annotations on a scale of 1-5.



(a) Evaluation on *sta_max_point_ca* task

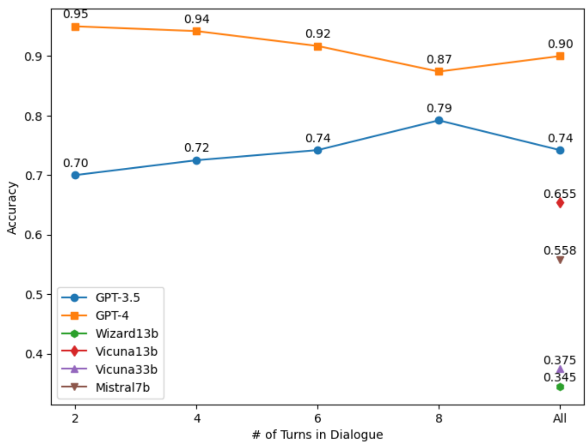


(b) Evaluation on *sta_max_point_dnd* task

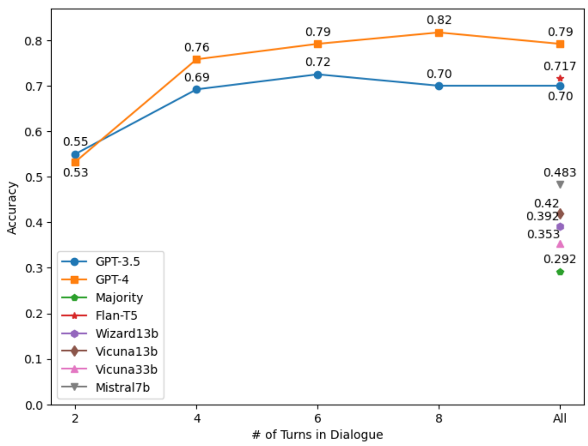


(c) Evaluation on *end_deal_total_ca* task

Figure 9: Evaluation on tasks demonstrating the impact of CoT prompting. We only consider GPT-4 for this analysis.



(a) Evaluation on *dur_ask_high_priority_ca* task



(b) Evaluation on *dur_partner_ask_high_priority_ca* task

Figure 10: Evaluation of two tasks not included in Figure 6, illustrating the impact of increasing the number of utterances seen by GPT-3.5 and GPT-4 models.

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	YOU: hello THEM: Hello, there. Should we discuss how to divide up the items for the camping trip? YOU: yeah. i need 2 firewood, 1 water and 1 food 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.		
	Response	Coherence	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 17: Sample responses generated by the models for an instance from the CaSiNo dataset. The *Coherence* and *Strategy* scores are based on human annotations on a scale of 1-5.