Evaluating Theory-of-Mind in Large Language Models Through Opponent Modeling

Kuru, Emre; Dogru, Anll; Dogan, Merve; Aydogan, Reyhan

DOI
10.1145/3717511.3747081

Publication date
2025

Document Version
Final published version

Published in
IVA 2025

Evaluating Theory-of-Mind in Large Language Models Through Opponent Modeling

Emre Kuru
Özyegin University
Istanbul, Turkiye
emre.kuru@ozu.edu.tr

Anıl Doğru
Özyegin University
Istanbul, Turkiye
anil.dogru@ozu.edu.tr

Merve Doğan
Özyegin University
Istanbul, Turkiye
merve.dogan@ozu.edu.tr

Reyhan Aydoğan
Artificial Intelligence and Data Engineering
Özyegin University
Istanbul, Turkiye
Interactive Intelligence Group
Delft University Of Technology
Delft, Netherlands
reyhan.aydogan@ozyegin.edu.tr

Abstract

Theory-of-Mind (ToM), the ability to infer the mental states, goals, and preferences of others — is a core component of human social intelligence. In this work, we investigate whether Large Language Models (LLMs) exhibit ToM capabilities in the context of strategic interaction. We frame opponent modeling in negotiation as a grounded and interpretable ToM task, where a model must infer an agent's preferences by observing offer exchanges during the negotiation. We guide LLMs to interpret offer histories and infer latent utility representations, including issue and value weights. We conduct a comprehensive evaluation of state-of-the-art LLMs across multiple negotiation domains. Our results show that LLMs can successfully recover opponents unknown preferences and in some cases even outperform classical opponent modeling baselines, even without task-specific training. These findings offer new evidence of LLMs' emerging capacity for social reasoning and position opponent modeling as a practical benchmark for evaluating Theory-of-Mind in foundation models.

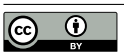
CCS Concepts

• **Computing methodologies** → *Intelligent agents.*

Keywords

Large Language Models, Negotiation, Theory of Mind

ACM Reference Format:
Emre Kuru, Anıl Doğru, Merve Doğan, and Reyhan Aydoğan. 2025. Evaluating Theory-of-Mind in Large Language Models Through Opponent Modeling. In *ACM International Conference on Intelligent Virtual Agents (IVA '25), September 16–19, 2025, Berlin, Germany.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3717511.3747081



IVA '25, Berlin, Germany

ACM ISBN 979-8-4007-1508-2/25/09
https://doi.org/10.1145/3717511.3747081

1 Introduction

Theory-of-Mind (ToM), the ability to infer the beliefs, desires, and preferences of others is a core component of human social intelligence [35, 43]. As Large Language Models (LLMs) become increasingly deployed in interactive settings, the question of whether they possess ToM-like reasoning has become central to understanding their capabilities and limitations. While recent studies have explored ToM in LLMs through tasks like belief attribution or psychological question answering, these approaches often rely on static, synthetic tests that fail to capture the complexities of dynamic interaction [3, 16, 46]. On this end, this study examines the ToM ability of LLMs in the task of *opponent preference modeling* (i.e., the task of inferring an agent's preferences from observed behavior in negotiation) by elaborately evaluating their performance in a grounded benchmark.

In multi-issue negotiations, an agent's preferences are typically represented by a utility function defined over a set of issues (e.g., delivery time), where each issue is assigned a weight indicating its relative importance, and each value (e.g., one day delivery) of an issue is assigned a score reflecting the agent's desirability for that value. During a negotiation, stakeholders exchange offers without knowing each other's preferences to come up with an agreement on a particular matter (e.g., price, delivery time and guarantee conditions in e-commerce) [6, 13]. Opponent modeling unfolds through these back-and-forths in negotiation, where agents interprets the opponent's offers to get insights into the underlying preferences of their opponents.

An agent with ToM capabilities should be able to observe these patterns and infer what the other party values most. This kind of inference reflects one of the most fundamental challenges of Theory-of-Mind: reasoning about unknown preferences based on partially observable behavior. It is important to emphasize that this inference is not just a matter of pattern recognition, as offers unfold over time, early bids tend to expose an opponent's true ideals, whereas later bids reveal what they are willing to concede under time pressure. Correctly interpreting this temporal interplay (i.e., knowing *when* a change reflects a genuine shift in preference versus a tactical concession) demands reasoning about unknown preferences and trade-off strategies. Such mental-state attribution cannot be achieved by simple pattern recognition alone but requires higher-level reasoning about the underlying nature of negotiation strategies.

In the case of demonstrating a reasonable performance, it is envisioned that these LLMs can be utilized in human-agent negotiations. Therefore, this study mainly focuses on negotiations where the number of exchanged offers are limited as it is usual in human-human negotiations. Recall that human negotiators do not have any tendency to exchange more than 10 offers on an average [8, 17, 28]. That is also challenging to learn opponent's preferences in such a few interactions.

Accordingly, we propose a structured prompting framework in which an LLM predicts these hidden issue and value weights using only the opponent's offer history. The model is not fine-tuned on negotiation data and relies solely on schema constraints and natural language instructions to guide its reasoning. Our results show that state-of-the-art LLMs can recover these preferences with great accuracy, outperforming classical opponent modeling strategies in some domains. These findings offer new evidence for the emergent Theory-of-Mind capabilities of LLMs and demonstrate that negotiation-based opponent modeling provides a practical and cognitively meaningful benchmark for evaluating social reasoning in foundation models.

To summarize, our contributions are as follows: (i) proposing a framework for opponent preference modeling in negotiation, providing a benchmark for evaluating Theory-of-Mind in LLMs; (ii) systematically evaluating a range of state-of-the-art LLMs with a rich set of metrics in our proposed human-centric benchmark (i.e., under limited observation settings), and (iii) demonstrating that LLMs can reliably infer opponent preferences and outperform classical opponent modeling baselines without task-specific training or optimization.

The remainder of this paper is organized as follows. In Section 2, we provide the necessary background on negotiation, with a particular focus on opponent preference modeling. Section 3 introduces the related work in this area, including prior studies on LLM reasoning and negotiation. Section 4 presents our proposed framework for evaluating opponent modeling as a Theory-of-Mind task. Section 5 details our evaluation setup, metrics, and results across a range of negotiation domains and model types. Finally, Section 6 summarizes our findings and discusses future directions for using opponent modeling as a benchmark for social reasoning in LLMs.

2 Background

Automated negotiation systems provide a framework in which two parties (e.g., humans, intelligent agents, or robots) interact to reach a mutual agreement. Each party keeps its preferences private and decides whether to propose, accept, or reject an offer under uncertainty until a specified negotiation deadline. The goal is to maximize individual utility while reaching a consensus. To ensure a structured process, a negotiation protocol governs the interaction. The most commonly used protocol is the *Stacked Alternating Offers Protocol* [5, 7], where one party initiates the negotiation by proposing an offer, and the other party responds by either accepting the offer or making a counter-offer. This process continues iteratively until the parties reach an agreement or the deadline is met.

In multi-issue negotiation domains, preferences are typically represented using an *additive utility function*, as shown in Equation 1. Each issue in the negotiation has an associated weight (W_i), with the total weight summing to 1.0. For each issue, multiple value options are defined in the domain, and each value is assigned a utility score ($V_i \in [0.0, 1.0]$). This formulation enables agents to evaluate and compare offers based on their overall utility.

$$U(o) = \sum_i W_i \times V_i(o) \tag{1}$$

A negotiating agent must determine both what to offer and when to accept an offer. For this purpose, agents typically include components such as a *bidding strategy*, an *acceptance strategy*, and an *opponent model* [10]. The bidding strategy generates offers by targeting a utility score based on factors such as elapsed time or observed opponent behavior [4, 12, 19, 25, 30, 36, 37, 41]. The acceptance strategy determines whether an incoming offer should be accepted [11]. The opponent model analyzes the opponent's past behavior and offers to extract useful insights that can guide decision-making process. A sophisticated agent should adapt its strategy by modeling the opponent's preferences and behavioral patterns to reach high-quality agreements efficiently. Moreover, estimating opponent's preference is essential for understanding its negotiation behavior [23]. Therefore, this study focuses specifically on the development of opponent models that predict opponent preferences.

Opponent models generally rely on the history of received offers to estimate hidden preferences. The most widely adopted opponent models are frequency-based approaches [37, 39, 42]. These models are typically based on two assumptions: (i) the parties make offers they are willing to accept, and (ii) they tend to concede over time due to deadline pressure. Based on these assumptions, most models infer that opponents begin with highly preferred offers and gradually concede. The fundamental study of frequency-based approach is introduced by van Krimpen *et al.* [42]. They propose a frequency-based heuristic that increases the estimated weight of an issue if its value remains unchanged across consecutive offers. This reflects the assumption that parties have a tendency to concede over their preferred issues due to the deadline pressure. As a result, the amount of increase decreases over time, modeled as $n = 1 - t$, where t denotes normalized negotiation time. The update rule for the issue score S_i is given in Equation 2, and the estimated issue weight $\hat{W}_i$ is obtained by normalizing the scores, as shown in Equation 3.

$$S_i = S_i + n \tag{2}$$

$$\hat{W}_i = \frac{S_i}{\sum_i S_i} \tag{3}$$

Similarly, value utilities are estimated based on their frequency of occurrence. Each time a specific value appears, its score S_i^j is incremented, as defined in Equation 4. The estimated utility $\hat{V}_i^j$ of a value is then computed relative to the most frequently observed value, as shown in Equation 5.

$$S_i^j = S_i^j + 1 \tag{4}$$

$$\hat{V}_i^j = \frac{S_i^j}{\max_k S_i^k} \tag{5}$$

While simple and effective in some scenarios, the classical frequentist heuristic is highly sensitive to the number of observed offers and the consistency of concessions. To reduce this sensitivity to noise, Tunalı *et al.* propose a windowed frequentist model, called *Scientist*, that compares non-overlapping windows of offers instead

of individual offer pairs [39]. This approach introduces two control parameters, α and β to adjust the amount of update, and increases the issue score (S_i) only when an issue remains unchanged across two consecutive windows and a concession is detected. The update rule is given in Equation 6. The calculation of issue weights and value scores largely follows the classical approach. However, to reduce the impact of noise when the number of observations is low, a smoothing factor γ is introduced during value score (S_i^j) calculation, as shown in Equation 7. This method improves robustness and accuracy in agent-agent negotiations, where a large number of offers are typically exchanged.

$$S_i = S_i + \alpha \times (1-t)^{\beta} \quad (6) \qquad \hat{V}_i^j = \left(\frac{S_i^j}{\max_k S_i^k}\right)^{\gamma} \quad (7)$$

These two frequency-based approaches are well-suited for agent-agent negotiation settings and often demonstrate strong performance in such scenarios. Although they typically require a large number of observations to yield reliable results, they have also shown promising outcomes in human-centric negotiations, where only a limited number of offers are available [17, 28]. Therefore, this study employs the Frequentist [42] (i.e., classical frequentist) and Scientist [39] (i.e., windowed frequentist) opponent models as benchmark methods for comparison.

3 Related Work

This section reviews related work on the Theory-of-Mind (ToM) abilities of large language models (LLMs), with a particular focus on studies that use negotiation as a testbed for evaluating ToM, as well as opponent modeling more in general, as it is our selected task to assess ToM in this study.

3.1 Theory-of-Mind in LLMs

ToM has emerged as a critical axis along which the social capabilities of LLMs are evaluated. Gandhi *et al.* introduce BigToM, a benchmark that probes different types of belief reasoning (e.g., forward and backward belief) using structured causal tasks [21] . Their findings show that models like GPT-4 perform near human-level on static inference tasks, suggesting that advanced LLMs exhibit proto-Theory-of-Mind behaviors. Zhou *et al.* argue that ToM evaluations must extend beyond inference to include action [45]. They propose Thinking-for-Doing (T4D), a framework where LLMs must act based on inferred mental states. Despite strong performance on traditional ToM tasks, models struggled to convert these inferences into successful decisions, especially in interactive contexts. Our work is complementary with these insights by embedding ToM evaluation in an interactive, decision-driven environment: negotiation. Rather than testing inference in isolation, we examine whether LLMs can reason about hidden preferences through ongoing interaction and apply these inferences strategically through structured opponent modeling.

Negotiation has emerged as a natural setting to assess the interactive reasoning, adaptability, and Theory-of-Mind (ToM) capabilities of LLMs. Foundational work by Gandhi *et al.* shows that prompted LLMs can simulate belief tracking and value estimation in matrix-game scenarios, revealing their potential for strategic reasoning [22]. Bakhtin *et al.* extend this idea to a more complex setting, demonstrating that hybrid LLM agents can achieve human-level performance in the Diplomacy game by combining dialogue-based belief inference with long-term strategic planning [14].

Building on these insights, recent studies have proposed comprehensive frameworks to evaluate a wide range of negotiation abilities in LLMs. Abdelnabi *et al.* design negotiation games to test adaptive behavior in both cooperative and adversarial settings [1]. Bianchi *et al.* present the *Negotiation Arena*, where LLMs engage in structured scenarios and apply tactics like simulated desperation and anchoring to improve performance [15]. Vaccaro *et al.* conduct a large-scale negotiation tournament involving over 120,000 rounds, revealing that warmth-oriented strategies increase deal rates and satisfaction, while manipulative behaviors expose vulnerabilities in LLM behavior [40].

Kwon *et al.* present a broad benchmark of 35 negotiation tasks including partner modeling, showing that while LLMs like GPT-4 perform well in objective partner modeling, they struggle with subjective inference (e.g., estimating satisfaction or trust) [32]. Instead of learning the opponent complete ordinal preferences, they aim to investigate whether the model correctly identifies the top-ranked issue. This simplifies the inherently ordinal nature of utility structures and does not capture the full distribution of preferences across all issues and values. In contrast, our work treats opponent modeling as a structured prediction task, requiring the model to generate complete utility functions in the form of normalized issue and value weights. This enables a richer and more cognitively complex evaluation of Theory-of-Mind capabilities, as it demands that LLMs recover fine-grained preference representations from observed negotiation behavior.

Complementing these structural frameworks, a parallel line of work focuses on the emotional and expressive aspects of negotiation. Yongsatianchot *et al.* show that combining verbal and non-verbal cues—such as facial expressions—leads to more socially effective negotiators [44]. Lin *et al.* measure the ability of large language models in identifying the content of the dialogue given in natural language (Whether the received dialogue is an offer or a preference statement) [33].

Together, these studies highlight the utility of negotiation for studying social reasoning. However, most focus on either dialogue generation for negotiation or general strategic behavior. Our work complements these by analyzing the offer exchanges to extract the opponents preferences in the form of additive utility functions commonly referred to as opponent preference modeling, a cognitively grounded sub-skill that reflects the core ToM challenge of inferring what the other party values based on offer dynamics, a crucial aspect that needs to be better understood and incorporated into LLM-based negotiation strategies.

3.2 Opponent Modeling in Negotiation

Accurately estimating an opponent's preferences is essential in negotiation settings, as each party typically has access only to its own preferences. Such estimation is critical for understanding the opponent's strategy, expectations, and behavior throughout the negotiation process. Several existing opponent models [24, 39, 42] demonstrate promising performance when a large number of observations are available. However, human-agent negotiations

are generally shorter, resulting in fewer observations and reduced model accuracy [8, 17].

To address this challenge, Keskin *et al.* introduce a conflict-based model [28], which extends the assumptions of frequency-based approaches by positing that human negotiators tend to concede more regularly and systematically. This model assumes that observed offers follow a discernible concession pattern and employs a majority voting mechanism to infer the most likely preference ordering. While effective under regular behavior, the model may struggle with noisy concession patterns, particularly when negotiating with agentic or inconsistent opponents.

In human-agent negotiations, participants may also exchange explicit arguments to express their preferences. Nazari *et al.* combine the classical frequentist approach with sentiment analysis to develop an argument-based opponent model [34]. They propose three heuristic strategies tailored to human negotiation behavior: (i) the issue-ratio heuristic, which uses the sequence of offers to estimate preferences; (ii) the issue-sentiment heuristic, which leverages explicit preference assertions detected in dialogue; and (iii) the offer/sentiment heuristic, which integrates both sources by averaging their outputs. This approach estimates only issue weights, assuming ordinal and known value utilities. As a result, it cannot be directly applied to more complex domains where prediction of value utility must also be taken into account.

Similarly, Doğru *et al.* propose an argument-based opponent modeling framework by extending the windowed frequentist model [17]. In their approach, argument types are extracted from human-agent negotiation dialogues and used to incrementally refine preference estimates. Hence, this integration of argument information enhances estimation performance. Although the method achieves promising accuracy in natural language negotiations, it depends on a predefined taxonomy of argument categories and classifiers trained on domain-specific data, which limits its scalability across diverse negotiation domains.

While these studies enhance both the quantity and quality of information used for preference estimation, they continue to rely on heuristic assumptions rooted in traditional models. Consequently, effectively modeling human-centric negotiations calls for new paradigms that move beyond classical, agent-oriented, and heuristic-driven approaches.

4 Proposed Approach

Understanding others' preferences based on limited interaction is a hallmark of social intelligence and a defining challenge for Theory-of-Mind (ToM). Negotiation systems provide a rich testbed for this ability. As agents exchange offers over time, they implicitly reveal their priorities through the structure and evolution of their offers. These behavioral cues convey a dynamic and interpretable signal for inferring underlying preferences, making negotiation a practical and cognitively meaningful context for evaluating ToM in language models.

To formalize this idea, we design a structured prompting framework in which a Large Language Model (LLM) observes the offer history from one side of a negotiation and infers the hidden utility function of that party. This task, commonly referred to as *opponent modeling*, requires reasoning about unobserved mental states from observable behavior and aligns closely with core ToM principles. Our framework casts opponent modeling as a structured prediction problem. The LLM receives a sequence of offers, along with metadata such as time constraints, and outputs a structured representation of preferences consisting of the following components:

- **Issue Weights**: A normalized distribution over issues, indicating their relative importance to the opponent.
- **Value Utilities**: A distribution over values within each issue, reflecting the opponent's desirability for those values.

Modern LLMs are powerful but prone hallucinate when asked for complex such structured outputs. In tasks like opponent modeling, where the model must return a complete nested JSON of issue weights and value utilities, any missing key or malformed field can break downstream parsing and evaluation. To prevent that, we utilize Python's Pydantic[1] library. Pydantic is a framework for defining data models with Python classes and type hints. Whenever you parse JSON through a Pydantic model, it automatically checks that every field is present, has the correct type, and meets any additional constraints—raising an error if something is missing or malformed. At runtime, we dynamically generate a Pydantic model that mirrors the current negotiation domain's issues and their allowable values. We insert its JSON schema into the LLM prompt and, once the model emits JSON, we parse it back through that same Pydantic model. Any validation failure (e.g. a missing/hallucinated issue or value) is caught immediately, which forces the LLM to produce only fully correctly formatted outputs; greatly improving the reliability of our opponent-modeling results. An example generated domain schema can be seen in Figure 1. After generating the domain schema, we embed it and a corresponding offer history into the prompt template below, before submitting it to the LLM.

```
{

  "IssueWeights": {
    "Accommodation": "number",
    "Destination":   "number"
  },
  "required": ["Accommodation", "Destination"],

  "ValueUtilities": {

    "Accommodation": {
      "Caravan": "number",
      "Hostel":  "number",
      "Hotel":   "number",
      "required": ["Caravan", "Hostel", "Hotel"]
    },

    "Destination": {
      "Berlin":   "number",
      "Paris": "number",
      "London":   "number",
      "required": ["Berlin", "Paris", "London"]
    }
  }
  "required": ["Accommodation", "Destination"],

"required": ["IssueWeights", "ValueUtilities"]

}
```

Figure 1: Example Pydantic Schema

[1] https://pypi.org/project/pydantic/

Prompt Template

You are an expert negotiator analyzing an opponent's preferences based on their offer history.
Your task is to identify patterns to determine their true preferences.

Given:
- A history of received offers
- Remaining time left when each offer was made
- The domain schema

Determine:
1. Issue weights of the opponent
2. Value weights per issue

Notes:
- Early offers reflect true ideals; later ones show concessions
- Frequent changes mean lower issue importance
- Infrequent changes mean higher importance
- Timing of changes reveals their strategy

Constraints:
- Sum of all issue weights must be 1.0
- For all issues, each value utility must be between [0,1]

Respond strictly in the given domain schema.

5 Evaluation

The performance of the proposed LLM-based opponent models is evaluated using the *NegoLog* platform [18], an open-source, Python-based automated negotiation framework designed for standardized agent evaluation and preference estimation. NegoLog provides structured logging of negotiation transcripts between autonomous agents across a wide range of pre-defined domains and enables the independent evaluation of opponent models by supplying negotiation histories, decoupled from specific agent strategies. In this study, the NegoLog framework is employed to incorporate LLM-based opponent models equipped with the proposed dynamic prompting mechanism. This integration enables structured preference inference directly from negotiation behavior. The platform simulates diverse negotiation interactions and generates offer histories, which serve as input to the LLM-based opponent modeling pipeline.

5.1 Experiment Setting

Since our focus is on Theory-of-Mind in human-centric settings, we select five multi-issue domains that have been extensively used in human–agent negotiation studies. Specifically, two from the Automated Negotiating Agents Competition (ANAC) [26] (*Car* [20], *Energy* [6]) and three from prior human–agent experiments (*Holiday* [17], *Fruit*, *Island* [29]).

Table 1: Negotiation Domain Characteristics

Domain	#Offers	Issues	Opposition	Balance Score
Car	240	[4,4,5,3]	0.22	−0.04
Holiday	256	[4,4,4,4]	0.28	0.00
Island	256	[2×8]	0.56	0.00
Energy	625	[5,5,5,5]	0.31	−0.02
Fruit	625	[5,5,5,5]	0.28	0.00

We chose these domains to cover a range of outcome-space sizes, issue complexities, and preference dynamics. These domains with their respective characteristics are described in detail in Table 1.

- **#Offers**: total number of possible agreements (reflects overall complexity).
- **Issues**: number of negotiable attributes per domain.
- **Opposition**: average conflict level, computed as the mean absolute difference between agents' utility functions over all outcomes. Higher values indicate more adversarial preferences, which stress test an opponent model's ability to discriminate.
- **Balance Score**: measures utility symmetry, defined as the mean utility difference normalized by outcome range. Values near zero denote balanced domains (equal potential gain), while larger magnitudes indicate inherent bias, challenging models to handle asymmetric trade-offs.

Similarly, the negotiation traces used in the evaluation are generated through simulations between agents employing human-like concession strategies. In particular, we employ four different negotiating agents, each of which reflects behavioral patterns commonly observed in human negotiators. *ConcederAgent* [41] is a time-dependent negotiator that begins with its most preferred offer and gradually concedes over time, mimicking typical human concession behavior. *ParsCatAgent* [6] also follows a time-dependent strategy but switches between several predefined concession tactics as time progresses, while introducing mild randomness in the offer selection process. *ParsAgent* [31] utilizes a frequency-based opponent model to investigate mutually acceptable offers, while following a time-dependent strategy and introducing mild randomness in the offer generation process. Finally, *HybridAgent* [30], which is specifically designed for human-agent negotiation, integrates time-based and behavior-based strategies within a unified decision framework.

In our evaluation, each agent negotiates with every other agent on each profile within the selected domains, with no repetitions, resulting in 60 negotiation sessions ($\binom{4}{2} \cdot 5 = 60$), where 4 and 5 denotes the number of agents and domains respectively. The deadline for each session is 20 rounds, which is generally sufficient for human negotiators [8, 17, 28]. For each offer in these negotiation sessions, every opponent-modeling strategy takes the opponent's bid history up to that round and outputs updated issue weights and value utilities. We apply this procedure twice per negotiation session, first using agent B's offer history to model agent B's preferences from agent A's perspective, and vice versa, yielding 120 inference tasks in total. Figure 2 illustrates this opponent modeling process from the perspective of an agent, performed independently of the agent's own internal strategy. In our simulations, 23 of the 60 negotiation sessions could not reach an agreement within the given deadline (20 rounds per session). When the agent were able to come to an agreement the negotiation sessions lasted on average 10.4 rounds. Providing a rich test-bed for the opponent modeling strategies through various negotiation scenarios.

Finally, we assess eight state-of-the-art LLMs: GPT-4o, O4 Mini, DeepSeek R1, Grok 3, LLaMA 4 Maverick, Claude 3.7 Sonnet, Gemini 2.0 Flash, and Gemini 2.5 Pro, selected for their strong reasoning performance and support for chain-of-thought prompting. Their

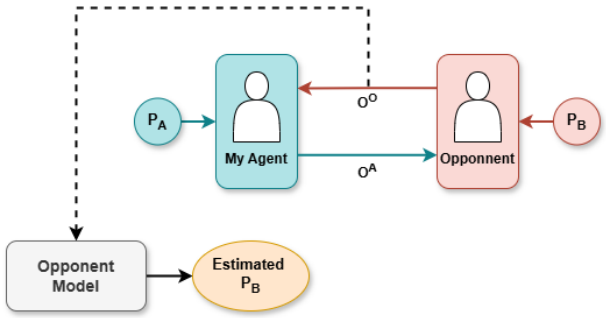


Figure 2: Opponent Model Work Flow

results are compared against two most commonly used opponent modelling baselines, Frequentist [42] and Scientist [39].

5.2 Evaluation Metrics

To evaluate the performance of an opponent model, the estimated preferences must be compared to the actual preferences. For this purpose, each offer in the negotiation domain is assigned a utility score based on both the estimated and actual preferences. The discrepancy between these two utility scores across all offers can be quantified using the *root mean squared error* (RMSE), a common metric in regression tasks [2], as shown in Equation 8. RMSE measures the prediction error in utility estimation, where lower values indicate better performance.

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{i}^{N}(U(o_i) - \hat{U}(o_i))^2} \tag{8}$$

In practice, accurately predicting the exact utility values of offers can be difficult due to the limited observations. However, for decision-making purposes, correctly identifying the relative preference between offers is often sufficient [12, 17, 23, 28]. Therefore, the ranking performance of the estimated preferences can be evaluated using rank correlation metrics [9]. Specifically, offers are sorted by both their estimated and actual utilities, and the correlation between these rankings is measured using *Spearman's rank correlation* [38] and *Kendall's Tau-b coefficient* [27].

To estimate correlation, potential outcomes are organized based on the learned opponent model, and this ranking is then compared with the actual ordering. As a result, the rank correlation (e.g., Spearman or Kendall's Tau) is computed between the actual outcome ranking and the estimated ranking. A high correlation indicates that both orderings are closely aligned. The correlation coefficient, denoted as **r**, ranges from -1 to 1, where the sign of the coefficient indicates the direction of the relationship, while its magnitude reflects the strength of the relationship. In other words, these metrics assess the consistency of the estimated ordering with the ground truth, where higher values indicate stronger agreement.

5.3 Results

Table 2 shows the average correlation results at the end of each negotiation session for each opponent modeling strategy. From these results we can observe that, Large Language Models (LLMs) generally outperform traditional baselines (e.g., *Frequentist* and *Scientist*) in most negotiation domains, particularly those with that are balanced or small-sized domains with moderate opposition, such as *Car* and *Holiday*. However, traditional models still show superior performance in more rigid high opposition domains such as *Island*.

Our results suggest that LLMs excel in domains where preference estimation depends on subtle patterns of consistency, concession, and trade-offs—hallmarks of strategic reasoning that reflect human negotiation. Domains like *Car* and *Holiday* feature evenly weighted issue spaces and moderate opposition, allowing LLMs to leverage their general reasoning abilities and schema-based prompt understanding. In contrast, traditional methods appear better suited for domains such as *Island*, where utility is often concentrated in a few high-priority issues, and opponent's behavior is more formulaic. In such settings, we observe that frequency-based heuristics can outperform general-purpose reasoning of LLMs by directly capturing dominant value signals through offer repetition. Overall, these results reveal that while LLMs are adept at capturing nuanced, context-dependent patterns, they still struggle in very large, or skewed domains where strong preferences dominate and clear statistical regularities can be exploited more effectively.

Table 2: Spearman and Kendall's Tau Correlations Across Domains and Models

	Spearman					Kendall's Tau				
Domain	Car	Fruit	Energy	Holiday	Island	Car	Fruit	Energy	Holiday	Island
Gemini 2.0 Flash	0.76 ± 0.07	**0.83 ± 0.07**	0.68 ± 0.24	0.78 ± 0.08	0.70 ± 0.16	0.58 ± 0.07	**0.65 ± 0.08**	0.53 ± 0.17	0.61 ± 0.09	0.54 ± 0.13
GPT-4o	0.71 ± 0.06	0.81 ± 0.05	0.67 ± 0.17	**0.80 ± 0.04**	0.60 ± 0.29	0.52 ± 0.06	0.61 ± 0.04	0.51 ± 0.13	**0.62 ± 0.04**	0.45 ± 0.22
Deepseek R1	0.71 ± 0.10	0.71 ± 0.13	0.56 ± 0.17	0.72 ± 0.05	0.65 ± 0.31	0.53 ± 0.08	0.54 ± 0.12	0.42 ± 0.13	0.53 ± 0.05	0.51 ± 0.25
Claude 3.7 Sonnet	0.73 ± 0.14	0.81 ± 0.08	0.67 ± 0.16	0.73 ± 0.10	0.68 ± 0.12	0.56 ± 0.12	0.63 ± 0.09	0.51 ± 0.15	0.55 ± 0.10	0.51 ± 0.11
Frequentist	0.71 ± 0.06	0.77 ± 0.03	0.74 ± 0.05	0.76 ± 0.04	**0.90 ± 0.05**	0.52 ± 0.06	0.58 ± 0.03	0.56 ± 0.05	0.57 ± 0.04	**0.74 ± 0.07**
Grok 3	0.76 ± 0.05	0.77 ± 0.06	0.65 ± 0.18	0.74 ± 0.06	0.85 ± 0.08	0.58 ± 0.05	0.58 ± 0.07	0.50 ± 0.14	0.55 ± 0.06	0.67 ± 0.08
Llama 4 Maverick	0.77 ± 0.07	0.70 ± 0.08	0.69 ± 0.10	0.72 ± 0.06	0.58 ± 0.36	0.58 ± 0.07	0.52 ± 0.06	0.51 ± 0.09	0.54 ± 0.06	0.47 ± 0.29
Scientist	0.52 ± 0.13	0.83 ± 0.03	**0.76 ± 0.07**	0.63 ± 0.07	0.89 ± 0.06	0.37 ± 0.10	0.64 ± 0.04	0.58 ± 0.06	0.46 ± 0.06	0.73 ± 0.07
O4 Mini	**0.80 ± 0.06**	0.82 ± 0.05	**0.76 ± 0.14**	0.70 ± 0.12	0.70 ± 0.30	**0.61 ± 0.06**	0.64 ± 0.05	**0.60 ± 0.12**	0.53 ± 0.11	0.57 ± 0.24
Gemini 2.5 Pro	0.77 ± 0.08	0.75 ± 0.08	0.71 ± 0.12	0.75 ± 0.07	0.77 ± 0.25	0.58 ± 0.08	0.57 ± 0.07	0.54 ± 0.11	0.57 ± 0.08	0.63 ± 0.20

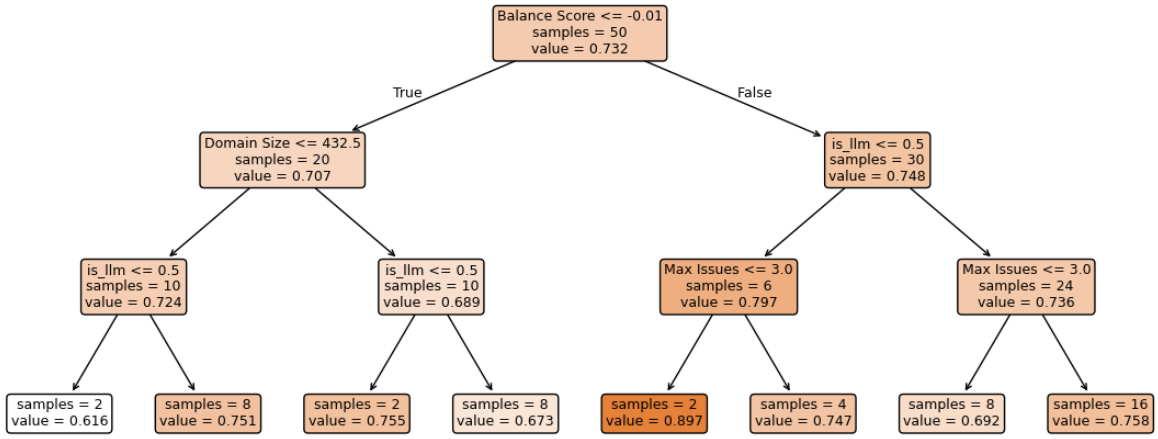


Figure 3: Decision Tree for Spearman Correlation

This pattern is further supported by the decision tree regressor constructed from our negotiation results as seen in Figure 3. Each instance in the collected negotiation dataset corresponds to a specific model evaluated within a given negotiation domain and is described by domain-level features such as domain size, number of issues, opposition, and balance score. We also include a binary feature, is_{llm}, which indicates whether the model is a LLM (is_{llm} = 1) or a traditional method such as Scientist or Frequentist (is_{llm} = 0). The tree is trained to predict the performance score of each model in its respective domain, enabling an interpretable analysis of how domain characteristics and model type affect performance. The decision tree reveals that LLMs outperform traditional methods in balanced domains. Furthermore, LLMs can also outperform traditional models in unbalanced domains when the domain size is small.

Furthermore, Table 3 presents the corresponding RMSE values across domains, offering a complementary perspective to the ranking-based metrics. While some domain related patterns persist (such as LLMs excelling in balanced or nuanced settings) the RMSE results reveal a critical distinction. Frequentist and Scientist models, despite their competitive ranking performance, consistently exhibit higher RMSE across most domains. This suggests that while traditional methods can approximate relative preferences effectively, they struggle to capture the precise magnitudes of utility, which may be useful for fine-grained decision-making. In contrast, LLMs like Claude 3.7 Sonnet and Gemini 2.0 Flash demonstrate stronger alignment between estimated and actual utilities.

Interestingly, the Spearman and RMSE results reveal complementary strengths across models (see Tables 2 and 3). Models like O4 Mini achieve top-tier Spearman's ρ (0.80 in *Car*) but exhibit only moderate RMSE (0.23 in *Car*), indicating they are excellent at ordering offers by preference yet less precise in estimating exact utility values. In contrast, models like Claude 3.7 Sonnet attain some of the lowest RMSE scores (0.13 in *Car*) but rank significantly lower on Spearman's ρ, suggesting they capture magnitude more faithfully at the expense of perfect ranking consistency. This contrast underscores how different LLM architectures exhibit distinct Theory-of-Mind reasoning patterns, and highlights the need to choose a model whose strengths best match the specific requirements of the task at hand.

Table 3: RMSE Across Domains and Models

	RMSE				
Domain	**Car**	**Fruit**	**Energy**	**Holiday**	**Island**
Gemini 2.0 Flash	0.17 ± 0.04	**0.13 ± 0.02**	0.22 ± 0.03	**0.15 ± 0.04**	0.23 ± 0.04
GPT-4o	0.24 ± 0.06	0.15 ± 0.03	0.25 ± 0.05	0.21 ± 0.04	0.19 ± 0.05
Deepseek R1	0.30 ± 0.06	0.29 ± 0.06	0.39 ± 0.12	0.34 ± 0.07	0.15 ± 0.06
Claude 3.7 Sonnet	**0.13 ± 0.02**	**0.13 ± 0.02**	0.21 ± 0.03	0.16 ± 0.03	0.22 ± 0.05
Frequantist	0.24 ± 0.04	0.21 ± 0.04	0.31 ± 0.06	0.26 ± 0.03	0.17 ± 0.04
Grok 3	0.21 ± 0.02	0.15 ± 0.03	0.21 ± 0.04	0.22 ± 0.03	0.15 ± 0.02
Llama 4 Maverick	0.25 ± 0.04	0.22 ± 0.04	0.34 ± 0.09	0.29 ± 0.04	0.20 ± 0.07
Scientist	0.21 ± 0.02	0.18 ± 0.03	**0.13 ± 0.03**	0.18 ± 0.03	0.30 ± 0.08
O4 Mini	0.23 ± 0.06	0.16 ± 0.04	0.26 ± 0.06	0.26 ± 0.07	0.15 ± 0.05
Gemini 2.5 Pro	0.26 ± 0.03	0.22 ± 0.04	0.32 ± 0.09	0.30 ± 0.05	**0.12 ± 0.04**

6 Conclusion

This study investigates the Theory-of-Mind (ToM) capacities of Large Language Models (LLMs) through the task of opponent preference modeling in negotiation. By requiring models to infer hidden preferences from limited offer histories, our benchmark targets a fundamental challenge in social reasoning: understanding hidden intentions based solely on behavioral traces. Our results show that state-of-the-art LLMs can accurately recover opponent preferences and often outperform classical heuristics in balanced, moderately complex domains, highlighting their emerging ToM-like abilities.

However, LLMs do not dominate across all settings. In domains with high opposition, or strong statistical regularities, traditional frequency-based models still perform competitively or even outperform LLMs. These findings reinforce the need for our benchmark: it does not merely serve as a showcase for LLM strength, but also reveals the boundaries of their current reasoning capabilities. As such,

it offers a principled lens for understanding when and why ToM-like inference emerges—or fails to emerge—in foundation models.

In future work, we propose extending this framework beyond preference estimation toward end-to-end negotiation. Specifically, enabling LLMs to leverage opponent models dynamically within their negotiation strategy would bridge ToM reasoning and interactive behavior. Our benchmark thus provides not only an evaluation suite for social inference, but a foundation for developing agents capable of strategic adaptation, collaboration, and negotiation in human-aligned settings.

References

[1] Sahar Abdelnabi, Amr Gomaa, Sarath Sivaprasad, Lea Schönherr, and Mario Fritz. 2024. Cooperation, competition, and maliciousness: LLM-stakeholders interactive negotiation. *Advances in Neural Information Processing Systems* 37 (2024), 83548–83599.

[2] Ethem Alpaydin. 2010. *Introduction to Machine Learning* (2nd ed.). The MIT Press, London, England.

[3] Maryam Amirizaniani, Elias Martin, Maryna Sivachenko, Afra Mashhadi, and Chirag Shah. 2024. Do LLMs Exhibit Human-Like Reasoning? Evaluating Theory of Mind in LLMs for Open-Ended Responses. *arXiv preprint arXiv:2406.05659* (2024).

[4] Reyhan Aydoğan, Tim Baarslag, Katsuhide Fujita, Johnathan Mell, Jonathan Gratch, Dave de Jonge, and et al. 2019. Challenges and Main Results of the Automated Negotiating Agents Competition (ANAC) 2019. In *Multi-Agent Systems and Agreement Technologies*. Springer International Publishing, Macao, China, 366–381.

[5] Reyhan Aydoğan, David Festen, Koen V. Hindriks, and Catholijn M. Jonker. 2017. *Alternating Offers Protocols for Multilateral Negotiation*. Springer International Publishing, Cham, 153–167. doi:10.1007/978-3-319-51563-2_10

[6] Reyhan Aydoğan, Katsuhide Fujita, Tim Baarslag, Catholijn M. Jonker, and Takayuki Ito. 2021. ANAC 2017: Repeated Multilateral Negotiation League. In *Advances in Automated Negotiations*. Springer Singapore, Singapore, 101–115.

[7] Reyhan Aydoğan, Koen V. Hindriks, and Catholijn M. Jonker. 2014. Multilateral Mediated Negotiation Protocols with Feedback. In *Novel Insights in Agent-based Complex Automated Negotiation*. Springer Japan, Tokyo, 43–59. doi:10.1007/978-4-431-54758-7_3

[8] Reyhan Aydoğan, Onur Keskin, and Umut Çakan. 2022. Would You Imagine Yourself Negotiating With a Robot, Jennifer? Why Not? *IEEE Transactions on Human-Machine Systems* 52, 1 (2022), 41–51. doi:10.1109/THMS.2021.3121664

[9] Tim Baarslag, Mark J.C. Hendrikx, Koen V. Hindriks, and Catholijn M. Jonker. 2013. Predicting the Performance of Opponent Models in Automated Negotiation. In *Proceedings of the 2013 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT) (WI-IAT '13, Vol. 2)*. IEEE Computer Society, Washington, DC, USA, 59–66. doi:10.1109/WI-IAT.2013.91

[10] Tim Baarslag, Koen Hindriks, Mark Hendrikx, Alexander Dirkzwager, and Catholijn Jonker. 2014. *Decoupling Negotiating Agents to Explore the Space of Negotiation Strategies*. Springer Japan, Tokyo, 61–83. doi:10.1007/978-4-431-54758-7_4

[11] Tim Baarslag, Koen Hindriks, and Catholijn Jonker. 2013. *Acceptance Conditions in Automated Negotiation*. Springer Berlin Heidelberg, Berlin, Heidelberg, 95–111. doi:10.1007/978-3-642-30737-9_6

[12] Tim Baarslag, Koen Hindriks, and Catholijn Jonker. 2013. *A Tit for Tat Negotiation Strategy for Real-Time Bilateral Negotiations*. Springer Berlin Heidelberg, Berlin, Heidelberg, 229–233. doi:10.1007/978-3-642-30737-9_18

[13] Tim Baarslag, Michael Kaisers, Enrico Gerding, Catholijn M Jonker, and Jonathan Gratch. 2017. When will negotiation agents be able to represent us? The challenges and opportunities for autonomous negotiators. International Joint Conferences on Artificial Intelligence.

[14] Anton Bakhtin, Jack Gray, Suraj Nair, Adam Lerer Xu, Tom Milani Alfredo, Ledell Wu, Eric Xiong, C. Lawrence Zitnick, Adam Dudzik, Anjalie Bhattacharyya, Jason Park, Gwern Branwen, Jonathan Fair, David Krueger, Yisong Halpern, Guillaume Lample, Dhruv Batra, and Jason Weston. 2022. Human-level play in the game of Diplomacy by combining language models with strategic reasoning. *Science* 378, 6624 (2022), 1067–1074. doi:10.1126/science.ade9097

[15] Federico Bianchi, Patrick John Chia, Mert Yuksekgonul, Jacopo Tagliabue, Dan Jurafsky, and James Zou. 2024. How well can llms negotiate? negotiationarena platform and analysis. *arXiv preprint arXiv:2402.05863* (2024).

[16] Zhuang Chen, Jincenzi Wu, Jinfeng Zhou, Bosi Wen, Guanqun Bi, Gongyao Jiang, Yaru Cao, Mengting Hu, Yunghwei Lai, Zexuan Xiong, et al. 2024. Tombench: Benchmarking theory of mind in large language models. *arXiv preprint arXiv:2402.15052* (2024).

[17] Anıl Doğru, Mehmet Onur Keskin, and Reyhan Aydoğan. 2025. Taking into Account Opponent's Arguments in Human-Agent Negotiations. *ACM Trans. Interact. Intell. Syst.* 15, 1, Article 2 (Jan. 2025), 35 pages. doi:10.1145/3691643

[18] Anıl Doğru, Mehmet Onur Keskin, Catholijn M. Jonker, Tim Baarslag, and Reyhan Aydoğan. 2024. NegoLog: an integrated python-based automated negotiation framework with enhanced assessment components. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence* (Jeju, Korea) *(IJCAI '24)*. Article 998, 4 pages. doi:10.24963/ijcai.2024/998

[19] Peyman Faratin, Carles Sierra, and Nick R. Jennings. 1998. Negotiation decision functions for autonomous agents. *Robotics and Autonomous Systems* 24, 3 (1998), 159–182.

[20] Katsuhide Fujita, Reyhan Aydoğan, Tim Baarslag, Koen Hindriks, Takayuki Ito, and Catholijn Jonker. 2017. *The Sixth Automated Negotiating Agents Competition (ANAC 2015)*. Springer International Publishing, Cham, 139–151. doi:10.1007/978-3-319-51563-2_9

[21] Shivanshu Gandhi, Yuxin Lyu, Zijian Lin, David Dohan, Douwe Kiela, Tushar Khot, Antoine Bosselut, and Karthik Narasimhan. 2023. Understanding Social Reasoning in Language Models with Language Models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics. https://arxiv.org/abs/2310.02239

[22] Shivanshu Gandhi, Jason Wei, Zijian Lin, Tushar Khot, Xuezhi Liu, Hyung Won Chung, Eunkyung Chung, Brian Lester, Ed H. Chi, Romal Thoppilan, Quoc V. Le, Barret Zoph, Aarohi Srivastava, David Dohan, Dragomir Radev, Kathy Lee, Percy Liang, and Hung Le. 2023. Strategic Reasoning with Language Models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics. https://arxiv.org/abs/2310.06770

[23] Koen Hindriks, Catholijn Jonker, and Dmytro Tykhonov. 2011. Let's dans! An analytic framework of negotiation dynamics and strategies. *Web Intelligence and Agent Systems* 9 (01 2011), 319–335.

[24] Koen Hindriks and Dmytro Tykhonov. 2008. Opponent Modelling in Automated Multi-Issue Negotiation Using Bayesian Learning. In *Proceedings of the 7th International Joint Conference on Autonomous Agents and Multiagent Systems - Volume 1* (Estoril, Portugal) *(AAMAS '08)*. International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, 331–338.

[25] Dave de Jonge and Carles Sierra. 2015. NB3: a Multilateral Negotiation Algorithm for Large, Non-linear Agreement Spaces with Limited Time. *Autonomous Agents and Multi-Agent Systems* 29, 5 (2015), 896–942.

[26] Catholijn Jonker, Reyhan Aydogan, Tim Baarslag, Katsuhide Fujita, Takayuki Ito, and Koen Hindriks. 2017. Automated negotiating agents competition (ANAC). In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 31.

[27] M. G. Kendall. 1945. The Treatment of Ties in Ranking Problems. *Biometrika* 33, 3 (1945), 239–251.

[28] Mehmet Onur Keskin, Berk Buzcu, and Reyhan Aydoğan. 2023. Conflict-based negotiation strategy for human-agent negotiation. *Applied Intelligence* 53, 24 (01 Dec 2023), 29741–29757. doi:10.1007/s10489-023-05001-9

[29] Mehmet Onur Keskin, Berk Buzcu, Berkecan Koçyiğit, Umut Çakan, Anil Doğru, and Reyhan Aydoğan. 2024. NEGOTIATOR: a comprehensive framework for human-agent negotiation integrating preferences, interaction, and emotion. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence* (Jeju, Korea) *(IJCAI '24)*. Article 1012, 4 pages. doi:10.24963/ijcai.2024/1012

[30] Mehmet Onur Keskin, Umut Çakan, and Reyhan Aydoğan. 2021. Solver Agent: Towards Emotional and Opponent-Aware Agent for Human-Robot Negotiation. In *Proceedings of the AAMAS* (Virtual Event, United Kingdom) *(AAMAS '21)*. International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, 1557–1559.

[31] Zahra Khosravimehr and Faria Nassiri-Mofakham. 2017. *Pars Agent: Hybrid Time-Dependent, Random and Frequency-Based Bidding and Acceptance Strategies in Multilateral Negotiations*. Springer International Publishing, Cham, 175–183. doi:10.1007/978-3-319-51563-2_12

[32] Youngjae Kwon, Sanya Shah, Vikram Yadav, Sweta Mishra, Aditi Chaudhary, Varun Narayana, and Marilyn A. Walker. 2024. Are LLMs Effective Negotiators? Systematic Evaluation of the Multifaceted Capabilities of LLMs in Negotiation Dialogues. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*. Association for Computational Linguistics. https://arxiv.org/abs/2402.12931

[33] Zijian Lin, Galen Hale, and Jonathan Gratch. 2023. Toward a Better Understanding of the Emotional Dynamics of Negotiation with Large Language Models. In *Proceedings of the 24th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*. Association for Computational Linguistics, 116–128. https://aclanthology.org/2023.sigdial-1.11

[34] Zahra Nazari, Gale Lucas, and Jonathan Gratch. 2015. Opponent Modeling for Virtual Human Negotiators. 39–49. doi:10.1007/978-3-319-21996-7_4

[35] David Premack and Guy Woodruff. 1978. Does the chimpanzee have a theory of mind? *Behavioral and brain sciences* 1, 4 (1978), 515–526.

[36] Victor Sanchez-Anguix, Reyhan Aydoğan, Vicente Julian, and Catholijn Jonker. 2014. Unanimously acceptable agreements for negotiation teams in unpredictable domains. *Electronic Commerce Research and Applications* 13, 4 (2014), 243–265.

[37] Victor Sanchez-Anguix, Okan Tunalı, Reyhan Aydoğan, and Vicente Julian. 2021. Can Social Agents Efficiently Perform in Automated Negotiation? *Applied Sciences*

11, 13 (2021), 1–26. doi:10.3390/app11136022

[38] C. Spearman. 1987. The Proof and Measurement of Association between Two Things. *The American Journal of Psychology* 100, 3/4 (1987), 441–471.

[39] Okan Tunalı, Reyhan Aydoğan, and Victor Sanchez-Anguix. 2017. Rethinking Frequency Opponent Modeling in Automated Negotiation. In *PRIMA: Principles and Practice of Multi-Agent Systems*, Bo An, Ana Bazzan, João Leite, Serena Villata, and Leendert van der Torre (Eds.). Springer International Publishing, Cham, 263–279.

[40] Michelle Vaccaro, Michael Caoson, Harang Ju, Sinan Aral, and Jared R Curhan. 2025. Advancing AI Negotiations: New Theory and Evidence from a Large-Scale Autonomous Negotiations Competition. *arXiv preprint arXiv:2503.06416* (2025).

[41] Rustam M. Vahidov, Gregory E. Kersten, and Bo Yu. 2017. Human-Agent Negotiations: The Impact Agents' Concession Schedule and Task Complexity on Agreements. In *50th Hawaii International Conference on System Sciences*, Tung Bui (Ed.). Hawaii, USA, 1–9.

[42] Thijs van Krimpen, Daphne Looije, and Siamak Hajizadeh. 2013. *HardHeaded*. Springer Berlin Heidelberg, Berlin, Heidelberg, 223–227. doi:10.1007/978-3-642-30737-9_17

[43] Andrew Whiten and RW Byrne. 1991. *Natural theories of mind: Evolution, development and simulation of everyday mindreading*. B. Blackwell Oxford.

[44] Nutchanon Yongsatianchot, Tobias Thejll-Madsen, and Stacy Marsella. 2024. Exploring Theory of Mind in Large Language Models through Multimodal Negotiation. In *Proceedings of the 24th ACM International Conference on Intelligent Virtual Agents (IVA '24)*. ACM, Glasgow, United Kingdom, 1–9. doi:10.1145/3652988.3673960

[45] Ben Zhou, Huaiyuan Zhang, Chunting Si, Rafal Kocielnik, Sameer Singh, and Xiang Ren. 2023. How Far Are Large Language Models from Agents with Theory-of-Mind?. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics. https://arxiv.org/abs/2310.02234

[46] Wentao Zhu, Zhining Zhang, and Yizhou Wang. 2024. Language models represent beliefs of self and others. *arXiv preprint arXiv:2402.18496* (2024).