Theory of Mind in Prisoner's Dilemma with Small LLMs

Aylin Gunal[1] , **Baihan Lin**[2] , **Djallel Bouneffouf**[3] ,
[1]University of Michigan
[2]Icahn School of Medicine at Mount Sinai
[3]IBM
gunala@umich.edu, baihan.lin@mssm.edu, djallel.bouneffouf@ibm.com

Abstract

In this work, we host a tournament of games of iterative prisoner's dilemma between LLMs and classic prisoner's dilemma strategies, as well as employ Theory of Mind (ToM) prompting. While previous works have focused primarily on the performance of large models, highlighting the capabilities of GPT4 in particular, we focus our investigation on smaller, cost-effective models and whether they demonstrate emergent social reasoning. Our results indicate that for the LLaMA and Falcon families, including ToM can cause cooperative behavior to significantly decrease, while the Qwen family tends to remain trusting of their opponents, despite the detriment to its performance and its accuracy in predicting its opponents next move.

1 Introduction

The capability to predict the beliefs and intentions of others is commonly referred to as Theory of Mind (ToM). As large language models (LLMs) increasingly participate in tasks that require interactive reasoning, there has been a significant rise in research that explores whether and when LLMs are able to exhibit ToM. LLMs' abilities to anticipate the belief states and intentions of other agents, including human agents, can indicate the viability of these models to successfully participate in multi-agent collaborative efforts.

The field of game theory is convenient in that it offers structured, constrained frameworks through which to study such agent behavior. While prior work has primarily focused on large-scale models with hundreds of billions of parameters [Akata *et al.*, 2023], [Lorè and Heydari, 2023], [Phelps and Russell, 2023], [Xie *et al.*, 2024], our study focuses on whether relatively small LLMs (3–8B parameters) can demonstrate emergent social reasoning in strategy games. We are motivated by the practical need for small, cost-effective models that can support agentic behavior in interactive environments. If small LLMs are able to exhibit ToM for positive overall interactions with other agents, this this may indicate viability for deployment in real-world applications and industry, where computational efficiency is desirable.

In this work, we focus on the classic experiment in game theory, the Prisoner's Dilemma. This game—particularly when implemented as an iterative game—requires agents to model not only the outcomes of their own actions but also to accurately predict the actions of other agents, making it a practical framework through which to investigate the relationship between ToM and strategic reasoning among agents. We quantitatively compare outcomes between models playing prisoner's dilemma with explicit consideration for ToM and those without, as well as investigate whether or not cooperative or selfish behavior scales with model size. We find that including consideration for ToM in prompts does affect model behavior, in line with previous works with massively large models [Akata *et al.*, 2023], [Zhang *et al.*, 2024]. In smaller models, the inclusion of ToM can result in a dramatic shift towards selfish actions, whereas with larger models the results vary.

2 Related Work

Game theory has recently been a popular framework through which to examine emergent strategic behavior with consideration for other players' actions in LLMs. Through iterative games such as multi-round prisoner's dilemma, LLMs are demonstrated to be able to form strategies based off of intermediate gameplay histories and respond accordingly to the actions of other agents [Akata *et al.*, 2023], [Xie *et al.*, 2024], [Mao *et al.*, 2023]. However, these strategies are not always *cooperative*; when playing lengthy games of iterative prisoner's dilemma with LLaMA2-70b and LLaMA3-70b models, as well as GPT3.5 and GPT4, all models except for GPT4 were found to be more likely to defect and act in self-interest even though building trust and cooperating consistently would have optimized their accumulated points [Fontana *et al.*, 2024]. A number of previous works across different games and set-ups have highlighted the performance of GPT4 in particular as a highly effective player capable of long-term planning, when compared to other LLM agents [Xie *et al.*, 2024], [Ni *et al.*, 2024], [Akata *et al.*, 2023], [Li *et al.*, 2023], although this doesn't necessarily translate to more cooperative behavior [Akata *et al.*, 2023]. As an added dimension to studying LLM behavior through game theory, previous experiments have frequently included forms of prompt manipulation including initializing agents with personas, configured with traits such as demographics and careers [Xie *et al.*, 2024].

LLMs generally maintain an accurate understanding of the

current game state throughout multiple rounds, although this ability diminishes with small LLMs such as the LLaMA2-7b [Fontana *et al.*, 2024]. Techniques such as meta-prompting, or otherwise inquiring about different aspects of the current game state, have been demonstrably effective towards improving LLM performance [Li *et al.*, 2023], although it is possible for LLMs to think *too* far ahead and subsequently take a hit to their performance [Zhang *et al.*, 2024]. The authors of [Yim *et al.*, 2024] demonstrate that although a specialized reinforcement-learning model comfortably outperforms LLMs in the strategy game Guandan, LLMs can also dramatically improve performance through ToM-based prompting with information about the current action space, without costly additional fine-tuning.

A clear outcome of these previous works is that ToM can be used in prompt engineering to improve individual performance in multi-agent interactions. However, a common theme among these studies is that models tend to act in their own self-interest, especially when explicitly considering their opponent's potential next move. In our work, we focus on considerably smaller models and whether or not ToM-prompting has a similar effect. We additionally make direct comparisons between models of the same families and of different sizes.

3 Methodology

We implement the prisoner's dilemma as an iterative game, allowing for models to potentially strategize across multiple rounds and games. A single game consists of several rounds, in which each player is made aware of the current game's history thus far and prompted to make a decision for the next round. Each pair of players plays against one another twice for five games with ten rounds each; we refer to each full set of games amongst all possible pairs of players as a tournament.

We run full tournaments under two prompt configurations: inclusive of ToM predictions and non-inclusive. The prompts are designed to request first-order ToM predictions, i.e. what action each player anticipates their opponent will take. All prompts include the game rules, the game history up until the current round, and instructions for response formatting. The full prompts that we use and details on prompt iterations are available in Appendix A.

We use two sets of models for all games: a set of LLMs with approximately 3 billion parameters, and a set of LLMs with approximately 8 billion parameters. Each small model has a counterpart in the same model family in the large model set. There are two types of tournaments: a tournament of smaller models, and a tournament with the larger models. We use the following LLMs:

SMALL LLMS: LLaMA-3.2-3b [Touvron *et al.*, 2023] (denoted as LLaMA3b), Qwen-2.5-3b [Yang *et al.*, 2024] (denoted as Qwen3b), Falcon-3-3b [Team, 2024] (denoted as Falcon-3b).

LARGE LLMS: LLaMA-3.1-8b [Touvron *et al.*, 2023] (denoted as LLaMA8b), Qwen-2-7b [qwe, 2024], Falcon-7b[1] (denoted as Falcon7b).

[1]https://huggingface.co/tiiuae/falcon-7b-instruct

Because of computational resource limitations, the LLMs in the LARGE LLMS tournament only play against non-LLM participants.

3.1 Experimental Setup

Prisoner's dilemma describes a strategy game in which two participants can play one of two actions simultaneously: cooperate or defect, with the following rewards structure:

	$Cooperate$	$Defect$
$Cooperate$	$(3,3)$	$(5,0)$
$Defect$	$(0,5)$	$(1,1)$

The optimal strategy in an iterative game of prisoner's dilemma would be to balance short-term gain with long-term outcomes. Cooperation results in mutual benefits (3,3), but defection leads to a higher payoff for the defector if the other cooperates (5,0), and a lower payoff for both players if both defect (1,1). The dilemma arises because, while mutual cooperation is the most beneficial outcome for both players, each player has an incentive to defect to potentially maximize their individual payoff, which can lead to the suboptimal outcome where both defect.

An equilibrium strategy is one that ensures a player maximizes their own outcome regardless of their opponent's actions. For example, a player who always defects is able to guarantee a payoff of at least 1 (or 5, if the other player cooperates), and subsequently will never have a worse payoff than their opponent. However, this strategy doesn't take advantage of the potential long-term benefits of cooperation, which could lead to higher total payoffs for both players if they choose to consistently cooperate.

In addition to the LLMs that play in the tournament, we implement several baseline strategies:

- **Single-Strategy:** This involves one of two strategies—always cooperating ("cooperate") or always defecting ("defect").
- **Grim-Trigger:** The grim-trigger strategy (denoted as "gt") starts with cooperation, but if the opponent defects at any point, the player immediately begins defecting for the remainder of the game, regardless of any future actions by their opponent.
- **Tit-for-tat:** Tit-for-tat (denoted as "t4t") describes mirroring the opponent's previous action. The player initially cooperates.
- **Random:** We incorporate three variants of a random strategy, with the following probabilities for randomly electing to cooperate or to defect. These strategies add an element of randomness to the decision-making process, offering a contrast to the deterministic nature of the other strategies:
  - A 50/50 distribution between cooperating and defecting, giving each action equal probability ("random").
  - A 25/75 distribution, where defection is more likely than cooperation ("more defect").
  - A 75/25 distribution, favoring cooperation more often than defection ("more cooperate").

4 Results

In this section, we conduct several analyses to understand tournament-level, game-level, and round-level patterns of behavior. P(C) denotes the probability of cooperation of a specific player, or the number of times that player cooperated across all rounds divided by the total number of rounds.

4.1 High-Level Patterns

Summary statistics for the SMALL LLMS prisoner's dilemma tournament both with and without ToM prompting are presented in Table 1. Per player, we include the total points accumulated over the course of the tournament, the percentage of games won (i.e. the player defected while the other player cooperated), as well as the percentage of ties, and P(C). Of all strategies, grim-trigger wins the most points and always defecting wins the most games. In this context, strategies that are more likely to defect have more successful outcomes.

The same summary statistics are computed for ToM prompting, and surprisingly we observe some key differences in the outcomes. Although grim-trigger remains the dominant strategy for accumulating points over the tournament, Falcon3b switches its strategy from highly cooperative with a cooperation probability of 90.30%, to defecting at every round when prompted using ToM. Ultimately, this switch pays off; Falcon3b has the second best performance for total points, with a significant point increase from non-ToM prompting, and the best win rate across the tournament, with a win rate nearly thirty times greater from non-ToM prompting. LLaMA3b has the third best performance for total points accumulation, and an increase in its win rate from non-ToM prompting. Notably, Qwen3b sees a *decrease* in performance, and is outperformed by classic strategies; however, unlike the other LLMs in the tournament, Qwen3b maintains a relatively high likelihood of cooperation.

In order to gain a better understanding of how ToM affects decision-making across the game, we compute action probabilities per round. As exemplified by the models in the SMALL LLMS tournament in Figure 1, the tournament outcomes indicate that regardless of ToM-prompting, LLMs tend to start off more cooperative and begin to defect later into the game.

Since the LARGE LLMS only played each classic strategy and not one another, we only include the summary statistics for the LLMs themselves in Table 2. We observe a similar pattern between non-ToM and ToM prompting for LLaMA8b; its performance improves for both total cumulative points and win rate, and its probability for cooperation decreases. Upon closer examination, we find that LLaMA8b adopts a grim-trigger strategy in response to ToM-prompting (see Table 3). Notably, both Qwen7b and Falcon7b take hits to their performances—more significantly for Qwen7b, similar to its 3b counterpart—between non-ToM and ToM prompting, as they both increase their likelihoods for cooperating.

We additionally compute the probabilities of each model defecting conditioned on the outcome of the previous round (if both players cooperate, this is denoted as CC; if the target player cooperates and their opponent defects, this is denoted as CD, etc). These probabililities are presented in Table 3. The LLaMA models see increases in defecting under all circumstances, although notably LLaMA3b is the only model that has a nonzero rate for defecting, both for non-ToM and ToM prompting, even when both models cooperated in the previous round. Generally speaking, models tended to maintain cooperation if the previous round induced cooperation for both players, which is line with the strategies of human players in extended games of prisoner's dilemma [Romero and Rosokha, 2017].

4.2 ToM-Specific Insights

In addition to analyzing the effects of ToM prompting on different models' strategies, we study each model's ToM ability. Each model's ToM prediction accuracy overall, the accuracy of prediction of specific actions, and the percent frequency of its predictions, are available in Table 4. LLaMA3b overwhelmingly predicts that its opponent will defect, which may explain why it sticks to a strategy of defecting even if the outcomes so far have been that both models cooperate (see Table 3). Notably, Qwen3b has a high overall prediction accuracy—with similar high accuracies for predicting both action types—even with a relatively low win rate and fourth-place ranking in terms of total accumulated points in the tournament (see Table 1). This may be explained by the Qwen family of models' dedication towards a high probability of cooperation and building trust across games, across different tournaments.

Among the LARGE LLMS, Falcon7b is more likely to trust and believe that its opponents will cooperate with it. LLaMA8b has a slightly more balanced distribution for its predictions, and has a high accuracy for predicting cooperation; alongside having a grim-trigger strategy under ToM-prompting, this indicates a highly practical approach to the games and maintains cooperation with those who pursue it as well.

5 Discussion

Through our experiments, we demonstrate that ToM prompting can affect the decision-making strategies of small LLMs in the prisoner's dilemma, frequently resulting in models to act prudently and selfishly, indicating that when these models are explicitly prompted to consider other agents' actions, the models act much more distrustful of their opponents. While models like Falcon3b and LLaMA3b benefitted from adopting more competitive strategies, Qwen3b's robust tendency towards cooperation ultimately limited its individual success. The adoption of grim-trigger strategies, especially among larger models, suggests that ToM allows even relatively small LLMs to engage in more complex, conditional strategies, but that these models need additional work towards engaging in long-term cooperative behavior with other agents.

A Prompts

Each model plays each other model in a tournament twice; once as Player 1 and once as Player 2. Player ID refers to either 1 or 2, depending on the specific game. The game history is initialized with "Here is the game history so far:

Model	**Total Points**	**Win Rate**	**Tie-Coop Rate**	**Tie-Defect Rate**	**P(C)**
Without ToM					
llama3b	2421	25.0%	26.0%	39.1%	35.90%
qwen3b	2555	10.5%	59.9%	23.2%	66.20%
falcon3b	2081	1.6%	64.0%	8.1%	90.30%
cooperate	2112	0.0%	70.4%	0.0%	100.0%
defect	2564	39.1%	0.0%	60.9%	0.0%
more cooperate	2033	14.7%	39.5%	11.3%	74.0%
more defect	2371	32.8%	9.7%	44.0%	23.2%
gt	2605	11.7%	59.2%	24.4%	63.9%
t4t	2359	5.9%	60.8%	24.0%	70.1%
random	2034	22.8%	21.1%	26.1%	51.1%
With ToM					
llama3b	2210	30.1%	2.8%	62.1%	7.8%
qwen3b	2058	8.8%	39.7%	42.7%	48.5%
falcon3b	2228	30.7%	0.0%	69.3%	0.0%
cooperate	1671	0.0%	55.7%	0.0%	100.0%
defect	2172	29.3%	0.0%	70.7%	0.0%
more cooperate	1601	11.8%	29.1%	13.8%	74.4%
more defect	1951	25.2%	6.2%	50.5%	24.3%
gt	2259	12.1%	41.8%	40.0%	47.9%
t4t	2111	6.2%	48.6%	34.3%	59.5%
random	1750	19.1%	16.7%	29.4%	51.5%

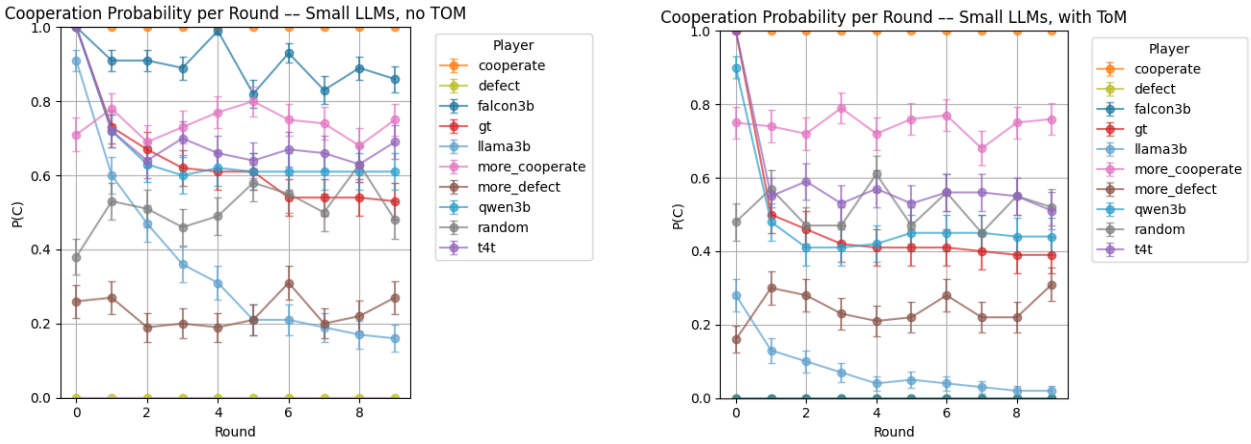
Table 1: High-level results for prisoner's dilemma tournament between SMALL LLMS and classic strategies, with and without ToM prompting.

Model	Total Points	Win Rate	Tie-Coop Rate	Tie-Defect Rate	P(C)
Without ToM					
llama8b	862	12.0%	52.0%	30.3%	57.7%
qwen7b	841	12.9%	52.6%	18.3%	68.9%
falcon7b	882	16.6%	45.7%	32.0%	51.4%
With ToM					
llama8b	889	15.4%	48.9%	30.3%	54.3%
qwen7b	686	9.4%	44.0%	16.9%	73.7%
falcon7b	875	13.7%	50.9%	28.9%	57.4%

Table 2: High-level results for prisoner's dilemma tournament between LARGE LLMS and classic strategies, with and without ToM prompting.

Model	*Without ToM*				*With ToM*			
	CC	CD	DC	DD	CC	CD	DC	DD
LLaMA3b	0.79	0.71	0	0.01	0.29	0.62	0	0.02
Qwen3b	1.00	0.28	0.02	0.01	1.00	0.37	0.06	0.01
Falcon3b	1.00	0.77	0.21	0.58	—	—	0	0
LLaMA8b	1.00	0.11	0	0	1.00	0	0	0
Qwen7b	1.00	0.59	0.08	0.12	1.00	0.95	0.35	0.04
Falcon7b	1.00	0	0	0	1.00	0.18	0	0

Table 3: Cooperation probabilities conditioned on the outcomes of the previous rounds. "—" indicates a previous outcome that never occurred in that player's games.



(a) Without ToM-prompting.

(b) With ToM-prompting.

Figure 1: Cooperation probabilities for SMALL LLMs by round.

Player	Accuracy	Coop. Pred. Acc.	Defect Pred. Acc.	Coop. Pred. Frequency	Defect Pred. Frequency
llama3b	0.68	0.62	0.68	0.026	0.974
qwen3b	0.81	0.80	0.82	0.490	0.510
falcon3b	0.73	0.55	0.82	0.338	0.662
llama8b	0.79	0.90	0.66	0.543	0.457
qwen7b	0.63	0.63	0.62	0.611	0.389
falcon7b	0.79	0.81	0.75	0.703	0.297

Table 4: ToM prediction accuracies and prediction distributions.

", and is concatenated with the results of each round thereafter in the format of "Round X: Player 1: Action 1, Player 2: Action 2" where each round is separated by a semicolon. We experiment with different formats and wording, and find that prompting for a single word as output is effective in the non-ToM prompts, and a list as output is effective in the ToM prompts.

System Prompt: "You are playing a game of iterative prisoner's dilemma with multiple rounds. Your goal is to maximize your points over the course of the entire game. In a round, if both players cooperate, both receive 3 points. If one defects and the other cooperates, the defector receives 5 points and the cooperator receives 0 points. If both players defect, both receive 1 point. Your vocabulary is restricted to 'Cooperate' and 'Defect'."

non-ToM User Prompt: "You are Player {playerid}. {gamehistory} Do you choose to 'Cooperate' or 'Defect' in the next round? Respond only with a single word."

ToM User Prompt: "You are Player {playerid}. {gamehistory} Do you choose to 'Cooperate' or 'Defect' in the next round? Additionally, what do you predict your opponent will choose to do in the next round? Respond with your own action and your predicted opponent's action in list format; for example, if you 'Cooperate' and you predict your opponent will 'Cooperate', respond: ['Cooperate', 'Cooperate']. Do not provide reasoning."

Ethical Statement

As far as we are aware, this study and its results present no ethical issues.

References

[Akata *et al.*, 2023] Elif Akata, Lion Schulz, Julian Coda-Forno, Seong Joon Oh, Matthias Bethge, and Eric Schulz. Playing repeated games with large language models. *ArXiv*, abs/2305.16867, 2023.

[Fontana *et al.*, 2024] Nicolo' Fontana, Francesco Pierri, and Luca Maria Aiello. Nicer than humans: How do large language models behave in the prisoner's dilemma? *ArXiv*, abs/2406.13605, 2024.

[Li *et al.*, 2023] Huao Li, Yu Quan Chong, Simon Stepputtis, Joseph Campbell, Dana Hughes, Michael Lewis, and Katia P. Sycara. Theory of mind for multi-agent collaboration via large language models. In *Conference on Empirical Methods in Natural Language Processing*, 2023.

[Lorè and Heydari, 2023] Nunzio Lorè and Babak Heydari. Strategic behavior of large language models: Game structure vs. contextual framing. *ArXiv*, abs/2309.05898, 2023.

[Mao *et al.*, 2023] Shaoguang Mao, Yuzhe Cai, Yan Xia, Wenshan Wu, Xun Wang, Fengyi Wang, Tao Ge, and Furu Wei. Alympics: Language agents meet game theory. *ArXiv*, abs/2311.03220, 2023.

[Ni *et al.*, 2024] Qin Ni, Yangze Yu, Yiming Ma, Xin Lin, Ciping Deng, Tingjiang Wei, and Mo Xuan. The social cognition ability evaluation of llms: A dynamic gamified assessment and hierarchical social learning measurement approach. *ACM Transactions on Intelligent Systems and Technology*, 2024.

[Phelps and Russell, 2023] Steve Phelps and Yvan I Russell. The machine psychology of cooperation: Can gpt models operationalise prompts for altruism, cooperation, competitiveness, and selfishness in economic games? *Journal of Physics: Complexity*, 2023.

[qwe, 2024] Qwen2 technical report. 2024.

[Romero and Rosokha, 2017] Julian Romero and Yaroslav Rosokha. Constructing strategies in the indefinitely repeated prisoner's dilemma game. *Game Theory & Bargaining Theory eJournal*, 2017.

[Team, 2024] Falcon-LLM Team. The falcon 3 family of open models, December 2024.

[Touvron *et al.*, 2023] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aur'elien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models. *ArXiv*, abs/2302.13971, 2023.

[Xie *et al.*, 2024] Chengxing Xie, Canyu Chen, Feiran Jia, Ziyu Ye, Kai Shu, Adel Bibi, Ziniu Hu, Philip H. S. Torr, Bernard Ghanem, and G. Li. Can large language model agents simulate human trust behaviors? *ArXiv*, abs/2402.04559, 2024.

[Yang *et al.*, 2024] Qwen An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxin Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yi-Chao Zhang, Yunyang Wan, Yuqi Liu, Zeyu Cui, Zhenru Zhang, Zihan Qiu, Shanghaoran Quan, and Zekun Wang. Qwen2.5 technical report. *ArXiv*, abs/2412.15115, 2024.

[Yim *et al.*, 2024] Yauwai Yim, Chunkit Chan, Tianyu Shi, Zheye Deng, Wei Fan, Tianshi ZHENG, and Yangqiu Song. Evaluating and enhancing llms agent based on theory of mind in guandan: A multi-player cooperative game under imperfect information. *ArXiv*, abs/2408.02559, 2024.

[Zhang *et al.*, 2024] Yadong Zhang, Shaoguang Mao, Tao Ge, Xun Wang, Yan Xia, Man Lan, and Furu Wei. K-level reasoning: Establishing higher order beliefs in large language models for strategic reasoning. 2024.