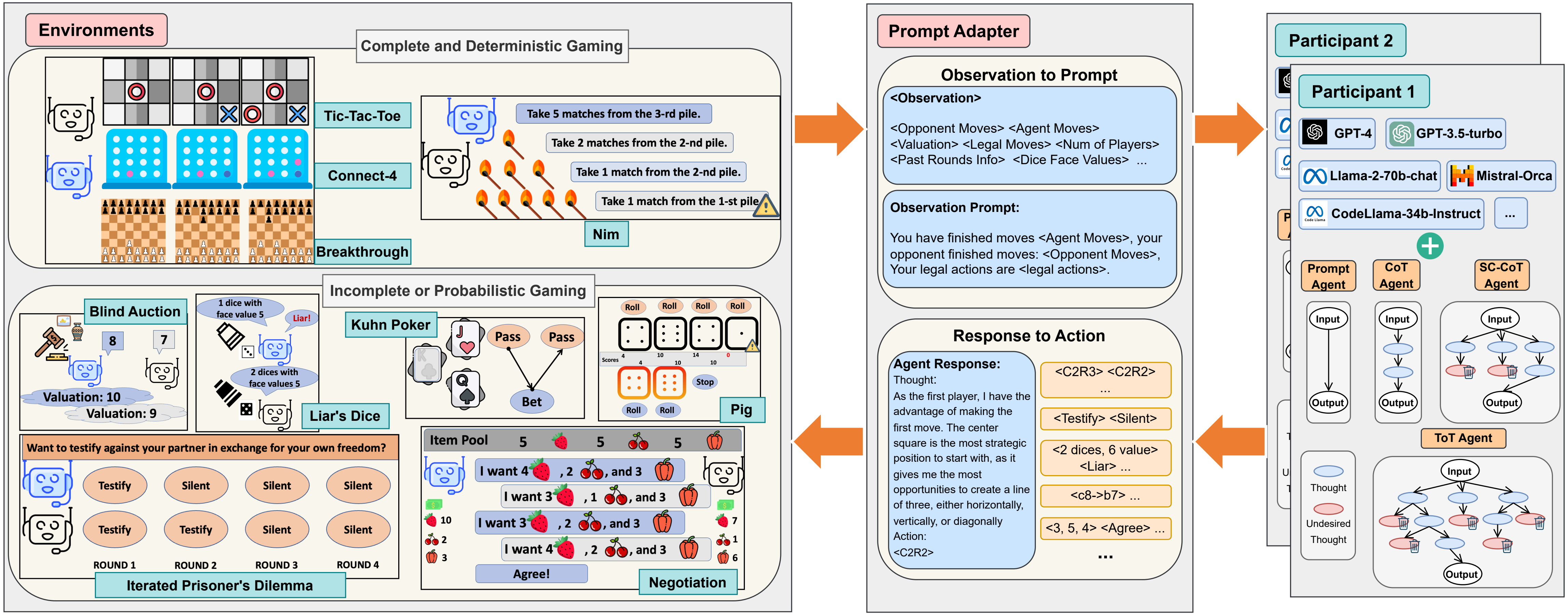
GTBench: Uncovering the Strategic Reasoning Limitations of LLMs via Game-Theoretic Evaluations

Jinhao Duan[1], Renming Zhang[2], James Diffenderfer[3], Bhavya Kailkhura[3], Lichao Sun[4], Elias Stengel-Eskin[5], Mohit Bansal[5], Tianlong Chen[5,6,7], Kaidi Xu[1]

[1]Drexel University [2]Boston University [3]LLNL [4]Lehigh University
[5]UNC Chapel Hill [6]MIT [7]Harvard University

HuggingFace: https://huggingface.co/spaces/GTBench/GTBench
Github: https://github.com/jinhaoduan/GTBench

LLM-vs-LLM for Reasoning Evaluation



- Rigorous rules and a well-defined action/state space, making them ideal for examining the strategic reasoning abilities of LLMs.

Introduction – Game Taxonomy, and Metrics

- Game Taxonomy

	Taxonomy of Games					Preferred Ability				
Game	Zero-Sum	First-player Advantage	▲ Complete ● Incomplete	▲ Dynamic ● Static	▲ Probabilistic ● Deterministic	Board Strategy	Bids	Collaboration	Bluff	Math
`Tic-Tac-Toe`	✓	✓	▲	●	●	✓	✗	✗	✗	✗
`Connect-4`	✓	✓	▲	●	●	✓	✗	✗	✗	✗
`Kuhn Poker`	✓	✓	●	●	▲	✗	✗	✗	✓	✓
`Breakthrough`	✓	✗†	▲	●	●	✓	✗	✗	✗	✗
`Liar's Dice`	✓	✗	●	●	▲	✗	✓	✗	✓	✓
`Blind Auction`	✗	✗	●	▲	▲	✗	✓	✗	✗	✓
`Negotiation`	✗	✗	●	●	▲	✗	✗	✓	✓	✓
`Nim`	✓	✓	▲	●	●	✗	✗	✗	✗	✓
`Pig`	✗	✗	▲	●	▲	✗	✗	✗	✗	✗
`Iterated Prisoner's Dilemma`	✗	✗	▲	▲	●	✗	✗	✓‡	✗	✓

† : Breakthrough has a slight first-player advantage which is not as significant as others.
‡ : The iterated version of Prisoner's Dilemma allows participants access to the actions made by their opponents in the past rounds, achieving implicit collaboration.

- Evaluation Metrics for LLM vs. LLM

Evaluation Metric: Normalized Relative Advantage. We introduce **Normalized Relative Advantage (NRA)**, denoted $\mathit{NRA}(\mathcal{M}_i, \mathcal{M}_o, f_s)$, to measure to relative advantage of $\mathcal{M}_i$ when competing against $\mathcal{M}_o$, under the score calculation f_s:

$$\mathit{NRA}(\mathcal{M}_i, \mathcal{M}_o, f_s) = \frac{\sum_m f_s(\mathcal{M}_i, m) - \sum_m f_s(\mathcal{M}_o, m)}{\sum_m f_s(\mathcal{M}_i, m) + \sum_m f_s(\mathcal{M}_o, m)},$$

Evaluation Metric: Elo Rating. Following the conventional rating mechanism in the real world, e.g., Chess, we employ the popular **Elo Rating** (Elo, 1960) for calculating the relative skill levels of players in zero-sum games. Please refer to Appendix A7 for more details of Elo rating.

Results – Various Game-Theoretic Scenarios

Complete and Deterministic Games

- LLMs always failed when competing against with optimal solver such as MCTS Agent

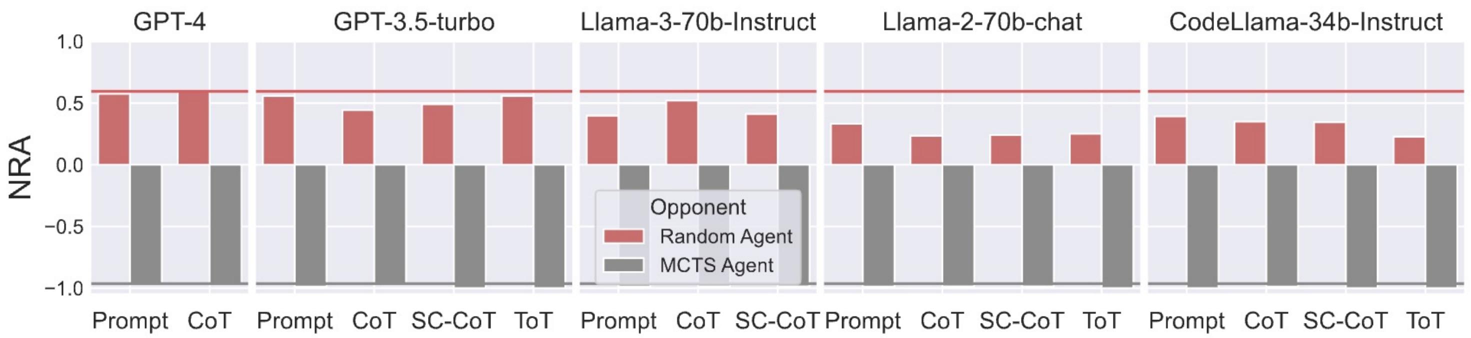


Figure 2: The NRA of state-of-the-art LLM-driven reasoning agents when against MCTS Agents and Random Agents, over complete and deterministic scenarios. Red and gray lines mean the maximum NRA achieved by LLM agents.

Incomplete and Probabilistic Scenarios

- LLMs achieves competitive performance compared with MCTS Agent in certain of games

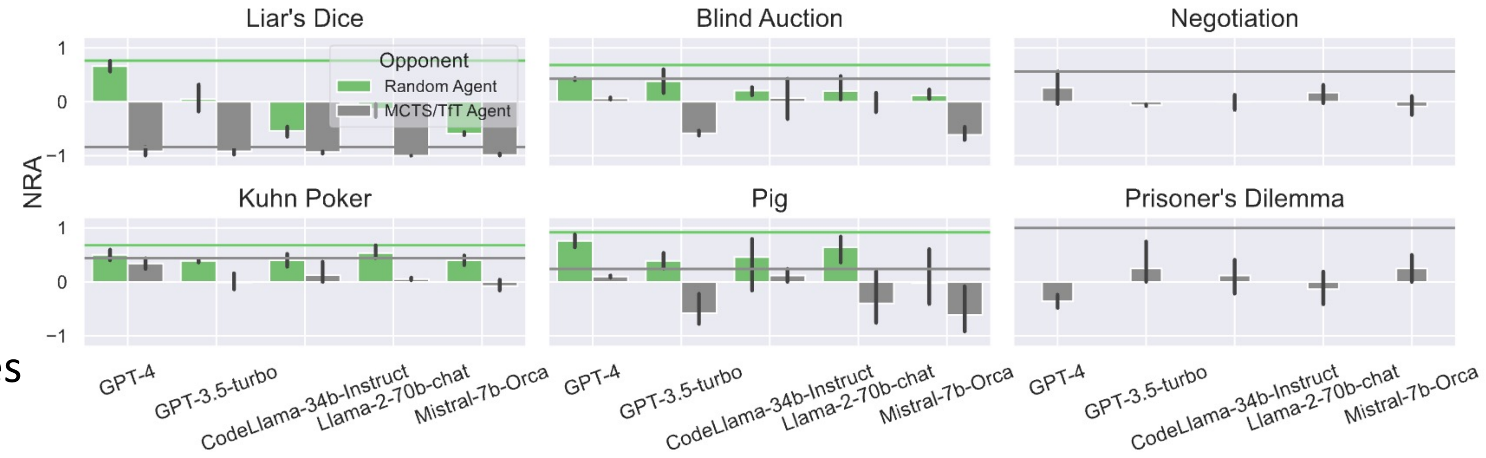


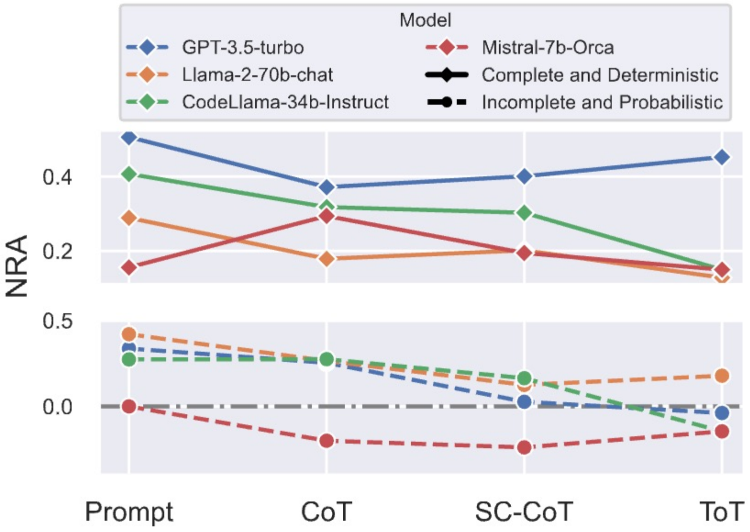
Figure 3: The game-wise NRA of LLMs when against MCTS/TfT Agents and Random Agents, over incomplete and probabilistic scenarios. Error bars are obtained over different reasoning methods. Green and gray lines mean the maximum NRA achieved by LLM agents.

Results – Advanced Reasoning and Code Pre-training Matters

Code Pre-training Benefits Strategic Reasoning

Model	avg. NRA in Det. Games	avg. NRA in Prob.	avg. NRA
GPT-4	0.09	0.15	0.13
Llama-3-70b-Instruct	-0.07	0.11	0.04
Llama-2-70b-chat	-0.25	-0.17	-0.20
CodeLlama-34b-Instruct	**-0.05**	**0.02**	**-0.01**
Deepseek-LLM-7b-chat	-0.09	-0.08	-0.08
Deepseek-LLM-67b-chat	**0.10**	-0.17	-0.05
Deepseek-Coder-6.7b-instruct	-0.14	**0.07**	**-0.03**

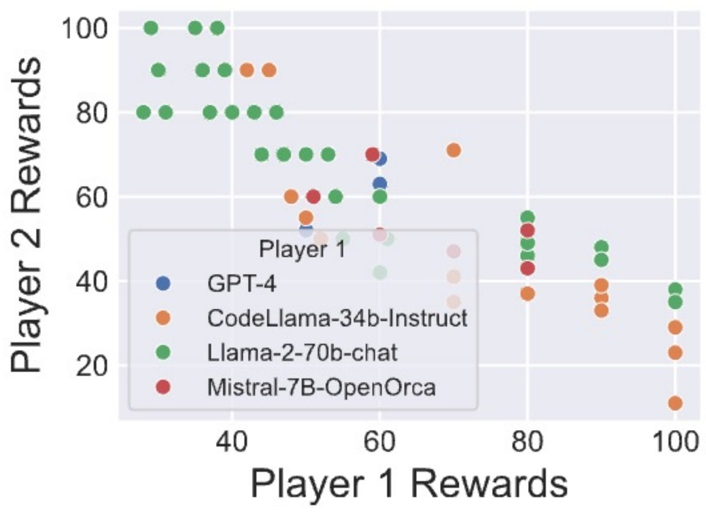
Advanced Reasoning Do Not Always Help



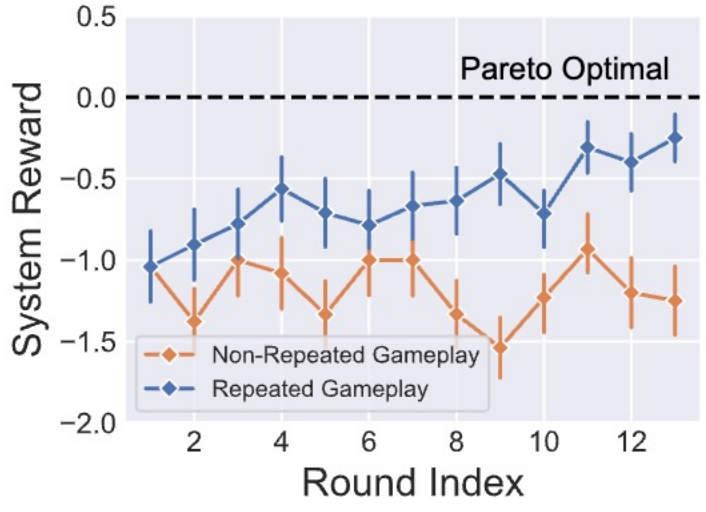
Opponent	Model	Reasoning	avg. NRA ↑
GPT-3.5-turbo w/ Prompt Agent	GPT-3.5-turbo	Prompt	0.00
		CoT	0.02
	Llama-3-70b-Instruct	Prompt	0.04
		CoT	0.07
	GPT-4	Prompt	0.13
		CoT	0.13
	CodeLlama-34b-Instruct	Prompt	-0.01
		CoT	-0.09
GPT-4 w/ Prompt Agent	CodeLlama-34b-Instruct	Prompt	-0.01
		CoT	-0.04
	Llama-2-70b-chat	Prompt	-0.10
		CoT	-0.23

Results – Game-Theoretic Properties

(a) Regret



(b) Resource Distribution



(c) Pareto Efficiency

GTBench: Uncovering the Strategic Reasoning Limitations of LLMs via Game-Theoretic Evaluations

HuggingFace: https://huggingface.co/spaces/GTBench/GTBench
Github: https://github.com/jinhaoduan/GTBench