
LLM Strategic Reasoning: Agentic Study through Behavioral Game Theory

Jingru Jia*, Zehua Yuan*, Junhao Pan, Paul E. McNamara, and Deming Chen

University of Illinois at Urbana-Champaign
{jingruj3, zehuay2, jpan22, mcnamar1, dchen}@illinois.edu

Abstract

What does it truly mean for a language model to “reason” strategically, and can scaling up alone guarantee intelligent, context-aware decisions? Strategic decision-making requires adaptive reasoning, where agents anticipate and respond to others’ actions under uncertainty. Yet, most evaluations of large language models (LLMs) for strategic decision-making often rely heavily on Nash Equilibrium (NE) benchmarks, overlook reasoning depth, and fail to reveal the mechanisms behind model behavior. To address this gap, we introduce a behavioral game-theoretic evaluation framework that disentangles intrinsic reasoning from contextual influence. Using this framework, we evaluate 22 state-of-the-art LLMs across diverse strategic scenarios. We find models like GPT-o3-mini, GPT-o1, and DeepSeek-R1 lead in reasoning depth. Through thinking chain analysis, we identify distinct reasoning styles—such as maximin or belief-based strategies—and show that longer reasoning chains do not consistently yield better decisions. Furthermore, embedding demographic personas reveals context-sensitive shifts: some models (e.g., GPT-4o, Claude-3-Opus) improve when assigned female identities, while others (e.g., Gemini 2.0) show diminished reasoning under minority sexuality personas. These findings underscore that technical sophistication alone is insufficient; alignment with ethical standards, human expectations, and situational nuance is essential for the responsible deployment of LLMs in interactive settings.

1 Introduction

The rapid development of large language models (LLMs) and generative AI has significantly broadened their applications, transitioning from basic text generation (41) and completion tasks to serving as sophisticated agents (45; 60; 21; 17). While existing benchmarks primarily assess LLMs on isolated language or math tasks (33; 23; 58), real-world deployment often demands more complex forms of decision-making, particularly strategic reasoning, where agents must interact with other entities whose actions directly influence the outcomes (64; 31; 22). Specifically, one-shot strategic reasoning refers to an agent’s ability to select a single, irreversible action where the outcome depends on both its own choice and that of other agents in this interaction. Consider an AI agent operating in a business environment where strategic interaction is critical. In procurement, for instance, it must anticipate supplier counteroffers when negotiating orders; in advertisement bidding, it predicts competitors’ bids to set optimal prices. In each case, the agent faces a one-shot decision that directly impacts its payoff and cannot be revised post-execution (43; 12).

Related Literature and Research Gap. Research on LLM decision-making typically begins by examining their behavior as independent decision-makers (4; 11; 34; 16). From an economic perspective, LLMs exhibit human-like patterns in preferences under uncertainty and risk (40). In social science, they have demonstrated alignment with human responses in moral judgment and

¹* Equal contribution

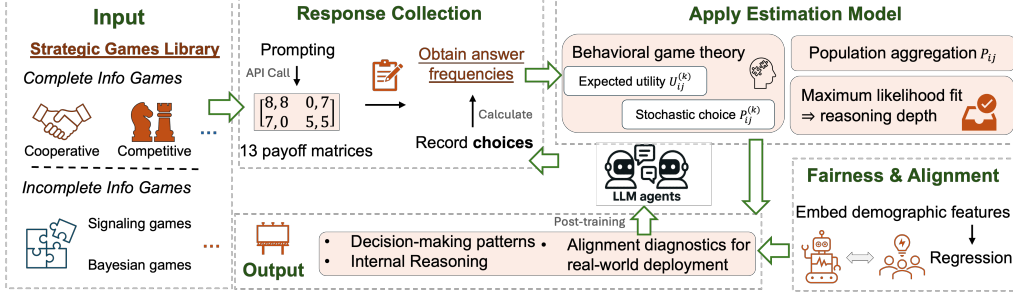


Figure 1: Framework overview

fairness-related tasks (15; 39; 52). Other studies further reveal alignments between LLMs and human judgments across diverse decision contexts (34; 46; 40). However, when demographic profiles are embedded into prompts, systematic inconsistencies and fairness concerns emerge (29; 40).

Strategic decision-making extends individual reasoning into interactive contexts, where outcomes depend not only on one’s own choices but also on the anticipated actions of others. Existing evaluations of LLM strategic reasoning are predominantly grounded in game theory. Several studies explore LLMs’ abilities in achieving rationality within game-theoretic frameworks, such as mixed-strategy Nash Equilibrium (NE) games (53) and classic strategic games like the prisoner’s dilemma (5; 28). Some works investigate the adaptability of LLMs to various prompts that influence cooperative and competitive behavior (50), and others evaluate their capacity to replicate human-like reasoning patterns in multi-agent settings (1; 61). While these studies provide a good starting point, many are limited to binary assessments of whether or not LLMs meet NE (35; 18) without quantifying their strategic reasoning capabilities or exploring the underlying mechanisms. While NE is a cornerstone concept in evaluating rational behavior, evaluating solely whether LLMs can achieve NE is an incomplete assessment of their reasoning capabilities (36; 32). The primary limitation of using NE as a measurement lies in its strict assumptions and focus on outcomes rather than the decision-making processes (3). It makes it difficult to interpret why LLMs deviate from optimal strategies. Also, NE assumes perfect rationality, which fails to account for the variability and bounded reasoning capability inherent in real-world problems. For LLMs—probabilistic models trained on human-generated data—this assumption is particularly problematic (63; 42). Their stochastic nature makes NE impractical as a comprehensive evaluation metric, and its assumption of perfect rationality falls into the circular reasoning fallacy when assessing LLMs’ rationality. Recent work begins to address these issues by incorporating stepwise evaluation (19) or applying models like Cognitive Hierarchy (CH) (65), but these approaches still rely on deterministic assumptions about context and agent type.

Key Contributions. To bridge this gap, we propose a framework, as shown in Figure 1, that moves beyond assessing LLMs with traditional game theory. We develop a method to quantify and characterize LLMs’ strategic decision-making, capturing bounded rationality, response to incentives, and belief formation. Our approach combines a diverse library of matrix games with a structured modeling framework based on Truncated Quantal Response Equilibrium (TQRE) stemming from behavioral game theory (51; 59). The key contributions are as follows:

1. We introduce an evaluation framework for assessing strategic reasoning in LLMs under various conditions. Applying to 22 state-of-the-art (SOTA) models on 13 abstracted real-world games, we find that massive models such as GPT-o3-mini and DeepSeek-R1 achieve top reasoning depth across most tasks, while smaller models occasionally match or even outperform them in specific game types.
2. Based on our evaluation results, we investigate why different models exhibit varying reasoning depth across games by analyzing the reasoning chains of three top-performing models in baseline competitive, cooperative, and mixed-motive games. Analysis reveals that each model’s performance is closely tied to its dominant reasoning style, but does not benefit well from long reasoning chains.
3. Finally, we examine the social effects and alignment of LLMs relative to human behavior when demographic features are assigned. We find that embedding demographic attributes prior to reasoning can reveal biases, even in advanced models such as DeepSeek-R1 and GPT-4o, particularly when processing minority-related features. This indicates that superior reasoning capabilities do not inherently lead to desirable or ethical outcomes, underscoring the need for careful calibration and a balanced approach in future LLM development.

2 Theoretical Foundations

2.1 Literature on Behavioral Game Theory

Classical game theory begins with the formalism of NE, in which subjects are assumed to be fully rational: they possess unlimited reasoning capacity, have perfect knowledge of the game, and make deterministic best responses (48). While mathematically sound, empirical evidence demonstrates frequent deviations from NE. For example, subjects often exhibit bounded rationality and stochastic behavior, such as overbidding in auctions or early cooperation in public goods games (25; 20). This gap between theory and observed behavior prompted the emergence of behavioral game theory, which progressively relaxes NE’s strong assumptions (7; 9; 8; 54).

The first core insight is that subjects do not always select the utility-maximizing option deterministically. Instead, they respond to incentives in a probabilistic pattern, leading to stochastic choice models, as proposed in the Quantal Response Equilibrium (QRE) model (44; 30). Another observation is that subjects differ in their depth of strategic reasoning levels. Rather than assuming that all players engage in unbounded thinking, models such as CH posit that individuals operate at varying levels of sophistication (54; 9). To capture both bounded reasoning and probabilistic behavior, we adopt the TQRE model, which combines the key behavioral refinements of QRE and CH. Instead of assuming convergence to a Nash fixed point, TQRE models subjects as reasoning at limited depths and responding stochastically, thus capturing more realistic strategic behavior, structured as below.

2.2 Modeling Preliminaries

Bounded-rational Belief Formation. Each subject draws a reasoning level $k \sim \text{Poisson}(\tau)$, $\tau > 0$, where τ is the average depth of strategic thought. A level- k subject does not know each opponent’s exact level $h < k$, so it mixes their level- h strategy distributions $\sigma_j^{(h)} : A_j \rightarrow [0, 1]$, $\sum_{a_j} \sigma_j^{(h)}(a_j) = 1$, using weights $W_h^{(k)}$, to form the weighted average of lower-level strategies. The marginal belief is

$$\bar{\sigma}_j^{(k)}(a_j) = \sum_{h=0}^{k-1} W_h^{(k)} \sigma_j^{(h)}(a_j).$$

Given that opponents act independently, the joint belief over all others’ actions $a_{-i} \in A_{-i}$ is

$$\Pr(a_{-i}) = \prod_{j \neq i} \bar{\sigma}_j^{(k)}(a_j)$$

, where the notation a_{-i} stands for the actions of all players except i .

Expected Utility (EU) and Stochastic Choice. Given $\Pr(a_{-i})$, a level- k subject’s expected payoff for each action $a_{ij} \in A_i$ is

$$U_{ij}^{(k)} = \sum_{a_{-i} \in A_{-i}} \Pr(a_{-i}) u_i(a_{ij}, a_{-i})$$

, where u_i is the payoff function. To allow for decision noise, set $\lambda_k = \gamma k$ with $\gamma > 0$, and convert utilities into choice probabilities:

$$p_{ij}^{(k)} = \frac{\exp(\lambda_k U_{ij}^{(k)})}{\sum_{a \in A_i} \exp(\lambda_k U_{ia}^{(k)})}.$$

Population-Level Aggregation. Finally, the overall probability that a randomly selected subject plays a_{ij} is the mixture over reasoning levels:

$$p_{ij} = \sum_{k=0} f_k p_{ij}^{(k)}, \quad f_k = \frac{\tau^k e^{-\tau}}{k!},$$

By estimating τ , we measure the reasoning depth LLMs entertain when making an interactive decision. We thus recover a cognitively grounded metric of bounded strategic reasoning that moves beyond NE-based assumptions to characterize how LLMs reason in multi-agent environments. For more details of the model specification, see the Appendix A.

Our modeling approach offers two key advantages over traditional NE-based evaluations. First, by incorporating bounded rationality and heterogeneous reasoning levels, it avoids the circularity of

assuming optimal behavior to test for rationality. Instead, it enables empirical estimation of strategic depth without presupposing NE play. Second, it captures the anticipatory nature of strategic reasoning: each agent’s EU is grounded in beliefs about opponents’ actual behavior, where τ quantifies the model’s depth of recursive belief modeling, i.e., layers of back-and-forth reasoning about others’ reasoning it engages in. This provides a cognitively meaningful measure of interactive sophistication.

Interpreting γ (decision precision). In the logit choice rule, γ scales how deterministically utilities are translated into actions conditional on beliefs. Larger γ yields more deterministic (near-best-response) behavior, while lower γ reflects noisier or more exploratory choice behavior. Importantly, γ does not encode preferences over outcomes—it captures the consistency of choice given utilities and beliefs. Because γ can be context- and scale-dependent, we report it primarily to avoid conflating shallow reasoning with stochastic execution, while using τ as our core measure of strategic sophistication. Full γ estimates are reported in Appendix B.

3 Experiment Setup and Estimation Framework

3.1 Game Library Design

We develop a library of 13 matrix games spanning 7 core types from behavioral game theory, grouped into complete-information (fully known payoffs) and incomplete-information (need to reason with uncertainty) settings. Each game varies in stakes and strategic structure.

Complete Information Games include four classic structures as the examples in Table 1. Each cell displays the payoffs for Player 1 (row chooser) and Player 2 (column chooser), respectively. Players choose their strategies, and the payoffs reflect the resulting outcome for each pair of chosen actions.

1. **Competitive:** Each player’s gain is the other’s loss, which tests adversarial reasoning.
2. **Cooperation:** Risky coordination, where mutual cooperation yields the highest joint payoff.
3. **Mixed-Motive:** Social dilemmas balancing private vs. joint outcomes (e.g., Prisoner’s Dilemma).
4. **Sequential:** temporal structure, with Player 1 moving first and Player 2 responding; we elicit decisions from the first mover.

Table 1: Representative matrices for complete information games

(a) Competitive	(b) Cooperation	(c) Mixed-Motive	(d) Sequential																										
<table><tr><td>10,-10</td><td>0,5</td><td>-5,8</td></tr><tr><td>-10,10</td><td>5,0</td><td>8,-5</td></tr><tr><td>0,0</td><td>5,-5</td><td>-5,5</td></tr></table>	10,-10	0,5	-5,8	-10,10	5,0	8,-5	0,0	5,-5	-5,5	<table><tr><td>8,8</td><td>0,7</td></tr><tr><td>7,0</td><td>5,5</td></tr></table>	8,8	0,7	7,0	5,5	<table><tr><td>3,3</td><td>0,5</td></tr><tr><td>5,0</td><td>1,1</td></tr></table>	3,3	0,5	5,0	1,1	<table><tr><td>0,5</td><td>0,3</td><td>0,0</td></tr><tr><td>5,2</td><td>3,3</td><td>-1,-1</td></tr><tr><td>2,4</td><td>4,3</td><td>0,-2</td></tr></table>	0,5	0,3	0,0	5,2	3,3	-1,-1	2,4	4,3	0,-2
10,-10	0,5	-5,8																											
-10,10	5,0	8,-5																											
0,0	5,-5	-5,5																											
8,8	0,7																												
7,0	5,5																												
3,3	0,5																												
5,0	1,1																												
0,5	0,3	0,0																											
5,2	3,3	-1,-1																											
2,4	4,3	0,-2																											

Incomplete Information Games incorporate belief-based reasoning about unknown payoff structures or player type. Examples of payoff matrices are shown in Table 2.

1. **Bayesian Coordination:** Players choose actions under uncertainty over the governing payoff matrix. Agents are given priors (e.g., 30% vs. 70%) over two types.
2. **Signaling:** The sender sees both real and fake matrices; the receiver must infer the true game. Payoffs reflect sender/receiver misalignment.

Table 2: Representative matrices for incomplete information games

(a) Bayes type 1	(b) Bayes type 2	(c) Signaling (sender)	(d) Signaling (receiver)																
<table><tr><td>10,10</td><td>5,2</td></tr><tr><td>7,5</td><td>3,3</td></tr></table>	10,10	5,2	7,5	3,3	<table><tr><td>8,8</td><td>6,3</td></tr><tr><td>5,4</td><td>2,2</td></tr></table>	8,8	6,3	5,4	2,2	<table><tr><td>5,5</td><td>2,1</td></tr><tr><td>3,2</td><td>1,0</td></tr></table>	5,5	2,1	3,2	1,0	<table><tr><td>4,4</td><td>6,3</td></tr><tr><td>2,3</td><td>1,2</td></tr></table>	4,4	6,3	2,3	1,2
10,10	5,2																		
7,5	3,3																		
8,8	6,3																		
5,4	2,2																		
5,5	2,1																		
3,2	1,0																		
4,4	6,3																		
2,3	1,2																		

We also include the SW10 matrix from (54), a benchmark for identifying human reasoning levels. The complete payoff matrix tables are attached in the Appendix. While our experiments use abstract normal-form games $(N, \{A_i\}, \{u_i\})$, these payoffs directly mirror real-world interactions. For example, in a recommender system, the e-commerce platform chooses to recommend one of two items, a_1 or a_2 , and the user choose to purchase (b_1) or not (b_2), yielding utilities $u_1(a_i, b_j)$ for the platform and $u_2(a_i, b_j)$ for the user. Similarly, in a negotiation, the proposer offers one of two deals, a_1 or a_2 , and the responder either accepts (b_1) or rejects (b_2), capturing the payoffs of agreement. Thus, payoff matrices capture essential strategic structures in various decision-making scenarios.

3.2 Evaluation Procedure

To quantify the average strategic reasoning depth (τ) from observed behaviors, we proceed as follows:

Step 1. Empirical Choice Frequencies. For each LLM and each game, we collect 30 independent trials and record the frequency c_{ij} with which the model chose action a_{ij} .

Step 2. Model-Implied Choice Probabilities. Under the bounded-rationality framework, a level- k agent’s probability of selecting a_{ij} is given by a logit function over its expected utility $U_{ij}^{(k)}$, scaled by $\lambda_k = \gamma \cdot k$. Aggregating across reasoning levels $k = 0, 1, 2, \dots$ with Poisson weight $f_k(\tau)$ yields:

$$p_{ij}(\tau, \gamma) = \sum_{k=0} f_k(\tau) \frac{\exp(\gamma k U_{ij}^{(k)})}{\sum_{a \in A_i} \exp(\gamma k U_{ia}^{(k)})}, \quad f_k(\tau) = \frac{\tau^k e^{-\tau}}{k!}.$$

Example: In a two-player game where each player $i \in \{1, 2\}$ has actions $A_i = \{a_{i1}, a_{i2}\}$ and Player 1’s utilities satisfy

$$u_1(a_{11}, a_{21}) = U_{11}, \quad u_1(a_{11}, a_{22}) = U_{12}, \quad u_1(a_{12}, a_{21}) = U_{21}, \quad u_1(a_{12}, a_{22}) = U_{22}$$

, one computes $U_{11}^{(k)}$ by summing each payoff U_{1j} weighted by the probability that the opponent (a level- $(< k)$ mixture) plays the corresponding action. That expected utility then enters the probability function to yield $p_{11}^{(k)}$, and the mixture over k produces $p_{11}(\tau, \gamma)$.

Step 3. Maximum Likelihood Fit. We estimate (τ, γ) by maximizing the log-likelihood of the observed counts under the model:

$$\max_{\tau, \gamma} \sum_{i,j} c_{ij} \ln p_{ij}(\tau, \gamma)$$

, where c_{ij} is the count of times action a_{ij} was chosen. Numerical optimization yields the values of τ that best fit the LLM’s choice distribution.

3.3 Model Selection.

Given the rapid evolution of LLMs in both open- and closed-source communities, we select representative models from each to demonstrate our framework’s effectiveness and highlight behavioral differences. While including more models is desirable, our selection captures common trends and supports meaningful evaluation. The selected models include GPT-4o (37), GPT-o1-preview (38), GPT-o3-mini (49), DeepSeek-R1 (27), DeepSeek-V2.5 (13), DeepSeek-V3 (14), gemma-2-27b-it (56), Gemini-2.0-Flash-Thinking (55), Granite-3.1-8B-Dense (47), Granite-3.1-3B-MoE, Claude-3-Opus (2), internlm2_5-20b-chat (6), Meta-LLaMA-3.1-405B-Instruct (26), Meta-LLaMA-3.1-8B-Instruct, QwQ-32B-Preview (57; 62), and glm-4-9b-chat (24).

3.4 Fit Diagnostics and Chance Baseline

We evaluate the maximum log-likelihood (MLL) for each model–game pair and contextualize these against a uniform **chance baseline**. For simultaneous games with m possible actions for Player 1 and n for Player 2, the baseline log-likelihood is defined as

$$\text{MLL}_{\text{chance}} = -\ln(mn).$$

For example,

$$\begin{aligned} 2 \times 2 &\Rightarrow \text{MLL}_{\text{chance}} = -\ln(4) = 1.386, \\ 3 \times 3 &\Rightarrow \text{MLL}_{\text{chance}} = -\ln(9) = 2.197. \end{aligned}$$

For sequential games, since only the first mover’s action is analyzed, the baseline reduces to

$$\text{MLL}_{\text{chance}} = -\ln(m),$$

where m is the number of available actions (e.g., $m = 3 \Rightarrow 1.099$). Values close to this baseline indicate random or level-0–like behavior, whereas higher (less negative) MLL values reflect better fit quality. Appendix B reports per-model fits and baselines, complementing Table 3.

4 Evaluation Results and Comparison

In this section, we present reasoning depth evaluations, followed by an analysis of reasoning variation across game types through models’ reasoning chains. By identifying distinct reasoning styles, we provide both quantitative and qualitative insights into how LLMs reason in interactive settings.

4.1 Strategic Reasoning Capabilities Across Models

Table 3 presents the reasoning depths evaluated through the framework for each model across games. Among all models, GPT-o1, GPT-o3-mini, and DeepSeek-R1 outperform others, with GPT-o1 ranking first in 6 games, GPT-o3-mini leading in 7 games, and DeepSeek-R1 securing the top position in 5 games. Some games have co-leaders, where multiple models achieve the highest evaluated strategic reasoning level. In these cases, certain LLMs consistently select strategies that align with theoretically robust principles while also demonstrating accurate anticipation of the opponent’s behavior. However, not all models achieve this level of capability, indicating variations in reasoning depth. This shows both the strengths of top models in structured settings and the effectiveness of our framework in distinguishing optimal decision-making from suboptimal or inconsistent reasoning. Additionally, our findings reveal significant variations in reasoning capabilities across different game types.

GPT-o1 consistently ranks among the top models in competitive and incomplete-information games, indicating a strong capacity for goal-oriented and adversarial reasoning. DeepSeek-R1 excels in cooperative and mixed-motive settings, likely due to its reinforcement learning-based optimization for social interaction and mutual benefit. GPT-o3-mini performs robustly across all game types, particularly in cooperative, mixed-motive, and incomplete-information scenarios, reflecting a balanced ability to manage trade-offs, adapt to uncertainty, and apply probabilistic inference.

The performance gap between stronger models, such as DeepSeek-R1, GPT-o1 and GPT-o3-mini, and weaker ones, like DeepSeek V2.5, DeepSeek V3, and GPT-4o, is notable across most games. Nevertheless, advanced models do not always dominate all game types. Smaller models like Gemma-2-27B and LLaMA-3.1-8B sometimes match or outperform larger counterparts, and DeepSeek-R1 is a top model despite not being the largest. Additionally, derived from a much smaller base model, R1-distilled models can outperform significantly larger models¹. This suggests that model size is not the sole factor in reasoning quality. Instead, context, model structure, and training data likely play a crucial role, highlighting the strong influence of contextual adaptation on LLMs’ reasoning abilities.

Table 3: Strategic reasoning depth (τ) across different game settings

	Complete Information Game									Incomplete Information Game			
	Competitive Games			Cooperation Games			Mixed-motive Games			Seq. Games	Bayesian Games		Signaling Games
	BL	HS	LS	BL	HP	AP	BL	H-Pun	L-Pun		p=0.5	p=0.9	
GPT-4o	1.543	1.936	0.729	0.602	0.143	1.505	1.665	0.537	1.499	0.222	1.973	1.953	3.590
GPT-o1	4.741	4.311	3.126	2.800	0.628	1.322	0.143	0.069	1.481	2.846	4.225	4.225	3.980
GPT-o3-mini	1.309	2.087	2.430	3.550	4.226	3.944	3.352	3.944	2.931	0.635	4.225	4.225	3.079
Gemma-V2	0.320	1.485	1.593	0.183	0.003	1.118	0.981	0.290	0.851	0.735	1.831	0.187	3.069
Gemini-2.0	1.202	1.202	1.958	2.487	1.609	0.070	2.461	0.223	1.230	1.683	4.225	4.225	3.920
Claude-3-Opus	1.131	0.982	0.036	3.390	1.322	1.505	1.333	1.000	1.220	0.717	2.014	4.225	3.623
Granite-3.1-8B	1.428	0.003	3.429	1.127	0.128	1.476	1.360	0.198	1.376	1.480	1.191	1.261	1.527
Granite-3.1-MoE	1.098	1.257	1.336	3.396	2.405	2.515	3.895	3.524	2.508	1.560	2.119	2.407	2.832
LLaMa-3.1-405B	0.988	0.910	0.177	0.379	0.001	1.118	1.368	0.290	1.180	0.688	1.658	0.226	1.543
LLaMa-3.1-8B	1.298	0.964	1.548	1.066	0.717	0.000	1.431	0.173	1.475	0.357	0.340	0.069	2.336
InternLM-V2	0.124	0.882	1.511	0.869	0.143	1.311	1.449	0.537	1.396	0.579	0.943	0.384	0.339
QwQ-32B	3.599	1.435	3.892	2.398	3.434	0.916	2.564	0.406	1.427	0.144	0.405	1.197	2.904
GLM-4	1.177	0.962	1.488	1.250	0.436	1.247	1.392	0.173	1.481	0.748	0.810	0.639	1.252
DeepSeek-V2.5	1.108	0.913	1.413	1.215	0.480	1.247	1.269	0.290	1.190	0.174	0.736	1.120	1.455
DeepSeek-V3	0.142	0.390	0.042	0.183	0.003	1.118	1.399	0.173	1.346	0.883	1.295	1.347	3.410
DeepSeek-R1	0.075	1.198	0.233	3.550	4.226	3.944	2.718	1.322	2.929	0.516	4.225	4.225	3.079

Note: **Bold values** indicate the highest strategic reasoning level (τ) observed for that game. Abbreviations: HS = High Stake, LS = Low Stake, BL = Baseline, HP = High Payoff, AP = Asymmetric Payoff, H-Pun = High Punishment, L-Pun = Low Punishment.

4.2 Empirical Exploration for Variations of Leading LLM in Different Game Types

To investigate why certain advanced models perform well in specific game types while others do not, we focus our analysis on three top-performing models, DeepSeek-R1, GPT-o1, and GPT-o3-mini, in baseline competitive, cooperative, and mixed-motive games.

¹Full results in Appendix B, Table 7

We gather available internal or summarized reasoning traces from each model and link them with its observed behavioral choices. For each model–game pair, we extract the empirical strategy distribution from 30 independent trials and identify the dominant or mixed-strategy pattern. We then examine the corresponding reasoning traces through qualitative analysis, aligning textual reasoning steps with revealed strategic behavior to evaluate coherence between what the model *says* and what it *does*. To ensure robustness and mitigate subjectivity, we collect a diverse set of reasoning traces from different models and anonymize them before analysis. Researchers who coded each CoT trace for reasoning style did so without knowing which model produced it, following a “blind review” procedure that minimizes post-hoc bias. A pre-defined structured coding rubric was used to classify reasoning styles (e.g., maximin, belief-based anticipation, opponent-oriented logic), and cross-validation was performed by having multiple researchers independently code overlapping subsets and compare results for consistency. Finally, we record the completion tokens required by each model to quantify computational demand and analyze whether longer reasoning chains correspond to improved or diminished decision quality.

Based on the summarization in Table 4, we observe that each model’s performance in different game types reflects the strategic logic it tends to apply. While these styles are not mutually exclusive or deterministic, they offer a coherent lens for interpreting variation in performance.

Table 4: LLMs’ reasoning styles and examples

Model	Dominant Reasoning Style	Representative Excerpts
GPT-o1	Consistent maximin, Payoff pessimism (risk-averse)	<i>“I select the row that maximizes our minimum payoff.” (competitive game)</i> <i>“Even though Row 0 gives 8 in the best case, Row 1 guarantees at least 5.” (coordinative game)</i> <i>“Worst-case payoff for me (assuming the opponent minimizes our payoff): 0” (mixed strategy game)</i>
GPT-o3	Belief-guided decision-making with fallback to maximin	<i>“Player 2 is likely to choose column 1, so I respond with row 0.” (competitive game)</i> <i>“Column 2 guarantees the highest minimum payoff... That’s why I selected column 2.” (mixed motive game)</i>
DeepSeek-R1	Opponent oriented decision, Lack of adversarial consideration	<i>“Player 2 will likely choose the column that maximizes their own payoff... so I choose row 0.” (competitive game)</i> <i>“... neither player benefits from deviating” (coordinative game)</i>

GPT-o1 consistently applies minimax reasoning, evaluating each option by its worst-case outcome. In nearly every reasoning chain we extracted, the model explicitly states, “I will now use minimax to help me make a decision.” This leads to strong performance in competitive games, where minimizing potential losses aligns with an optimal strategy. However, the same strategy becomes overly cautious in cooperative or mixed-motive games, where achieving mutual benefit often requires trust and risk-taking. GPT-o1 also exhibits a highly defensive stance, at times assuming that the opponent’s objective is to minimize its payoff, interpreting interactions as fully adversarial. GPT-o3-mini demonstrates greater flexibility. It primarily engages in belief-based anticipation, attempting to infer the opponent’s likely move and respond accordingly. This allows it to perform well in cooperative and mixed-motive settings, where alignment with the other player is key. Yet under uncertainty, GPT-o3-mini sometimes falls back to the worst-case logic, which can limit its potential in competitive games but also protects against major failures. DeepSeek-R1 adopts a strategy that begins with assumptions about the opponent’s likely action, typically based on the idea that the opponent will pursue their own payoff. This logic works well in cooperative games, where the players’ incentives are aligned, but lacks the adversarial caution required in competitive games. R1 also tends to exhibit strategic trust, assuming that when mutual benefit is unavailable, the opponent will not deviate simply to harm the other player, in contrast to GPT-o1, which at times assumes spiteful intent.

An additional observation relates to token usage, as shown in Appendix B Table 8. Contrary to intuitive expectations, higher token counts in internal reasoning do not correlate with better reasoning. In fact, GPT-o1 and DeepSeek-R1 are leaders in competitive and cooperative games, respectively, while they produce the shortest CoT outputs within their strongest games. We observe that longer reasoning chains often suggest hesitation and uncertainty, rather than deeper insight. Conversely, more concise CoT responses tend to signal higher confidence and more direct strategy choices. Notably, in competitive games, DeepSeek-R1 often exhibits repeated self-doubt in its CoT, but this results in redundant reasoning loops that inflate token usage without improvement. This points to future research on improving token efficiency in CoT generation, such as incorporating RL-based training that encourages decisive and diverse reasoning over repetitive patterns.

Table 5: CoT effect on reasoning examples

Model	Prompt Excerpt
Example of Improvement:	"I should pick the row that gives me the maximum value among these worst-case... Wait, I made a mistake... Let me reconsider using the concept of maximin..."
Claude-3-opus	"Row 0 and Row 2 both have a worst-case of -5, but Row 0 offers a potential payoff of 10, so I choose Row 0."
Example of Decline:	"We assume Player One picks the row that minimizes our payoff, so..."
GPT-4o	"Break tie with average... Column 2 is better."

4.3 Effect of CoT Prompting

Our empirical study of top-performing models with embedded internal CoT reasoning motivates us to explore whether CoT prompting can also enhance models that do not natively generate internal reasoning. The strategic reasoning level after explicitly adding the CoT prompt is presented in Appendix B Table 9. Intuitively, many believe CoT can likely enhance a model’s reasoning ability. However, our results show that CoT’s impact on strategic reasoning is mixed, with no consistent improvement across all models and game types, which is consistent with prior findings suggesting that the "CoT does not always help" (18; 10), particularly in strategic settings.

For example, Figure 2 compares τ in the Competitive-Base game with and without CoT prompting. While some models improve, others remain unchanged or even decline. We select two representative cases: with the largest improvement and sharpest decline: Claude-3-Opus and GPT-4o. Table 5 presents excerpts from their CoT reasoning chains to illustrate the underlying dynamics ².

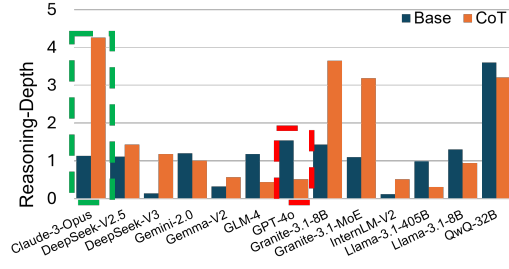


Figure 2: Comparison between baseline and CoT
Note: Green marks the largest increase; red marks the largest decline

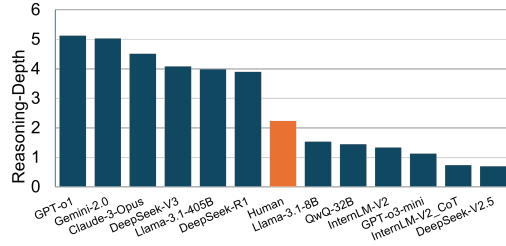


Figure 3: Comparison among human participants and LLMs on the SW game

If we look at their gaming strategies, Claude-3-Opus quickly identifies a clear attempt to apply the maximin strategy at an earlier stage, an appropriate approach for the competitive game. It initially makes calculation errors, referring to incorrect cells during its payoff analysis, but the model explicitly identifies and corrects these mistakes mid-chain. This self-correction enables it to recover and ultimately arrive at a robust decision that aligns with its strategic intent. In contrast, GPT-4o starts with a fundamentally flawed assumption, mistakenly believing that the opponent will act adversarially to minimize its payoff. Based on this incorrect premise, the model applies a tie-breaking rule using average values across columns, stuck in a verbose reasoning loop that reinforces the wrong strategy. These again imply that CoT does not guarantee better reasoning. When a model begins with a sound strategy, CoT can support useful self-correction and refinement. However, when the underlying reasoning is misaligned from the beginning, extended chains often lead to repetitive and unproductive loops rather than meaningful improvement, which echoes one of our findings in Section 4.2.

5 Detecting Social Effect with Human

While reasoning depth is useful for evaluating LLMs’ strategic capacity, model sophistication alone is not enough. Real-world deployment requires alignment with contextual goals, ethical constraints, and human norms. To examine these factors, we compare LLM reasoning with human performance in identical games and examine how assigned demographic personas influence model behavior, shedding light on identity-induced strategic shifts and fairness in AI decision-making.

²Full reasoning chains are in Appendix E

Figure 3 compares all models against human subjects in a single-scenario setting (54; 51), with six models surpassing human performance. However, this ranking should not imply that “a higher score is better”. In many applications, alignment with human reasoning is more critical than outperformance. Our framework supports such evaluation by quantifying reasoning depth across subjects, offering a path to calibrate LLMs for human-aligned behavior. For instance, businesses can use representative human samples to fine-tune models toward user-aligned decisions.

To further evaluate the alignment of LLMs with human strategic reasoning, we construct a diverse set of socio-demographic personas spanning a broad range of individual characteristics, inspired by (29). These personas are structured into two tiers: foundational demographic features (Panel 1) and advanced identity attributes (Panel 2), as detailed in Table 6.

Table 6: The Personas across 10 socio-demographic groups that we explore in this study.

Group	Persona
Panel 1: Foundational Demographic Features	
Sex	male, female
Education Level	below lower secondary, lower secondary, upper secondary, short-cycle tertiary, bachelor, and graduate degrees
Marital Status	never married, married, widowed, divorced
Living Area	rural, urban
Age	15 - 24, 25 - 34, 35 - 44, 45 - 54, 55 - 64, 65+
Panel 2: Advanced Demographic Features	
Sex Orientation	heterosexual, homosexual, bisexual, asexual
Disability	physically-disabled, able-bodied
Race	African, Hispanic, Asian, Caucasian
Religion	Jewish, Christian, Atheist, Other Religious
Political Affiliation	lifelong Democrat, lifelong Republican, Barack Obama supporter, Donald Trump supporter

We then perform Ordinary Least Squares (OLS) regressions using the estimated behavioral parameters as dependent variables. The general specification of the regression is:

$$\text{Reasoning_Depth}_i = \beta_0 + \beta_n D_{ni} + \epsilon_i$$

Here, i indexes each observation, D_{ni} is the n -th binary demographic indicator, β_n denotes its estimated effect on the behavioral parameter, and ϵ_i is the error term.

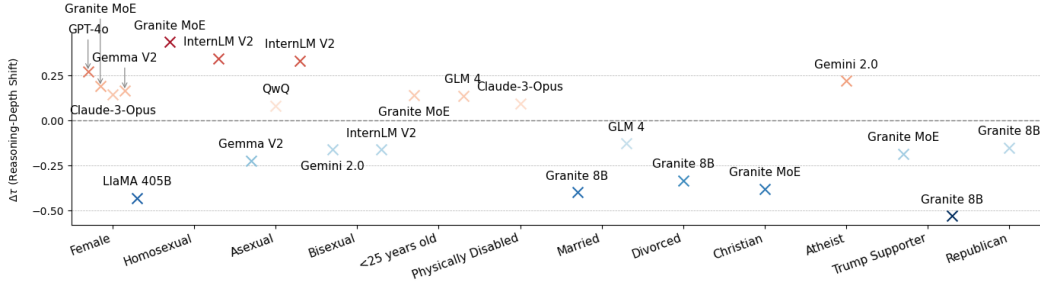


Figure 4: Significant demographic shifts in reasoning depth

Note: Each point reflects the significant change in reasoning depth relative to a reference group through OLS. Reference categories are: Male (gender), Heterosexual (sexual orientation), Age 25–64 (age), Able-bodied (disability status), Single (marital status), Other Religious (religion), and Democrat (political affiliation). Positive (negative) shifts indicate higher (lower) reasoning depth relative to the reference group.

Figure 4 shows estimated shifts in reasoning depth across demographic conditions with statistically significant effects from OLS regressions. A consistent pattern emerges around gender: when prompted with a “female” persona, several large models—such as GPT-4o, InternLM V2, and Claude 3-Opus—exhibit increased reasoning depth, suggesting deeper multi-level reasoning under female identity framing. Sexuality-based variation reveals further heterogeneity. While “homosexual” and “bisexual” personas raise reasoning depth in models like GPT-4o and GLM 4, they lead to significantly reduced τ in Gemini 2.0 and Granite MoE. These divergent effects highlight both the contextual flexibility of LLM reasoning and the risk of identity-induced attenuation. Other demographic factors can also cause disturbance in reasoning depths. We include full results of this study in Appendix B.

Mechanism note. Observed persona-induced shifts likely stem from statistical associations in training corpora that link demographic descriptors to behavioral patterns, potentially modulated by reinforcement learning from human feedback (RLHF). Models often deny explicit bias yet reveal

implicit adaptation under identical payoffs, indicating latent contextual priors. Our framework surfaces these implicit shifts and quantifies their magnitude for downstream auditing and fairness alignment.

6 Conclusion and Discussion

This study sets a new milestone in evaluating LLM strategic reasoning by assessing while accounting for contexts, addressing the limitations of NE-focused benchmarks. We find that strong performance on standard tasks (e.g., math) does not always guarantee higher reasoning depth. Analysis also reveals the limitations of CoT prompting in strategic settings, where its benefits are not universal and can, at times, impair reasoning. These findings inform ongoing efforts to refine LLM reasoning workflows for complex decision-making. Additionally, our demographic analysis exposes how biases can emerge even in top-performing models, underscoring the need for fairness-aware evaluation.

Effectiveness of CoT. While CoT can enhance the reasoning process in language models, its effectiveness hinges on task comprehension and alignment with intended objectives. Our analysis shows that well-aligned CoT prompts support structured reasoning and improved outcomes, whereas misaligned prompts can amplify errors—even in models trained with RL from Human Feedback (RLHF). DeepSeek-R1 illustrates that, despite strong performance in some games, it underperforms in others, exhibiting repetitive self-questioning and redundant token generation without better solutions. RLHF helps refine strategies, but without the intrinsic ability to evaluate context-dependent strategies, models may remain trapped in ineffective reasoning loops. These findings highlight the need for adaptive, context-aware reasoning and raise a key open question: how can external prompts be better aligned with a model’s internal reasoning processes to ensure robust and interpretable behavior?

Alignment and Fairness in AI Reasoning. Superior strategic reasoning depth does not equate to ethical decision-making. When demographic cues are introduced, models can adjust their strategies based on the given role, indicating that their decisions can encode and present societal biases. This is especially concerning in multi-agent interactions, where solely prioritizing strategic efficiency may lead to inequitable or adversarial outcomes. In interactive settings, LLMs that adjust strategies based on demographic cues risk perpetuating bias in negotiations and resource allocation, leading to real-world inequalities over time. This presents a critical challenge to balance strategic reasoning with fairness. Unconstrained optimization may encourage exploitative tactics, while rigid fairness constraints could hinder adaptability. Future development must move beyond performance metrics to integrate ethical safeguards, ensuring that LLMs engage in fair and justifiable decision-making in dynamic multi-agent environments.

Limitations and Future Directions: While our study sheds light on LLM reasoning patterns and benchmark alignment, the causal link between prompts and decision outcomes remains underexplored. Future work should investigate how individual cognitive skills—such as logic, math, and contextual understanding—interact to shape reasoning capabilities. Moreover, our current framework is semi-dynamic; extending it to real-time multi-agent interactions may reveal new reasoning dynamics and performance disparities. Exploring such settings could provide deeper insights into adaptive LLM behavior and strategy formation.

Broader Impact: This study provides practical insights into how the internal decision-making processes of large language models translate into real-world social outcomes through quantitative evaluations. Unlike explicit discrimination, these biases emerge through adaptive reasoning, making them more difficult to detect and regulate. As organizations increasingly rely on AI to guide complex decisions affecting diverse populations, researchers and practitioners must remain cautious: subtle reasoning biases in sophisticated AI could inadvertently widen existing societal gaps. Addressing these challenges today will ensure that LLMs become inclusive tools promoting societal equity, rather than unintentionally reinforcing historical injustices.

Acknowledgment

We gratefully thank the anonymous reviewers and area chair for their insightful comments and constructive feedback. We also acknowledge the support provided in part by the AMD Center of Excellence at the University of Illinois Urbana-Champaign.

References

- [1] Aher, G.V., Arriaga, R.I., Kalai, A.T.: Using large language models to simulate multiple humans and replicate human subject studies. In: International Conference on Machine Learning. pp. 337–371. PMLR (2023)
- [2] Anthropic: The claude 3 model family: Opus, sonnet, haiku (2024), <https://api.semanticscholar.org/CorpusID:268232499>
- [3] Baba, V.V., HakemZadeh, F.: Toward a theory of evidence based decision making. *Management decision* **50**(5), 832–867 (2012)
- [4] Brand, J., Israeli, A., Ngwe, D.: Using gpt for market research. Harvard Business School Marketing Unit Working Paper (2023)
- [5] Brookins, P., DeBacker, J.M.: Playing games with gpt: What can we learn about a large language model from canonical strategic games? Available at SSRN 4493398 (2023)
- [6] Cai, Z., Cao, M., et al.: Internlm2 technical report (2024)
- [7] Camerer, C.F.: Progress in behavioral game theory. *Journal of economic perspectives* **11**(4), 167–188 (1997)
- [8] Camerer, C.F.: Behavioral game theory: Experiments in strategic interaction. Princeton university press (2011)
- [9] Camerer, C.F., Ho, T.H., Chong, J.K.: A cognitive hierarchy model of games. *The Quarterly Journal of Economics* **119**(3), 861–898 (2004)
- [10] Carlander, D., Okada, K., Engström, H., Kurabayashi, S.: Controlled chain of thought: Eliciting role-play understanding in llm through prompts. In: 2024 IEEE Conference on Games (CoG). pp. 1–4. IEEE (2024)
- [11] Chen, Y., Liu, T.X., Shan, Y., Zhong, S.: The emergence of economic rationality of gpt. *Proceedings of the National Academy of Sciences* **120**(51), e2316205120 (2023)
- [12] Dalrymple, D., Skalse, J., Bengio, Y., Russell, S., Tegmark, M., Seshia, S., Omohundro, S., Szegedy, C., Goldhaber, B., Ammann, N., et al.: Towards guaranteed safe ai: A framework for ensuring robust and reliable ai systems. arXiv preprint arXiv:2405.06624 (2024)
- [13] DeepSeek-AI: Deepseek-v2: A strong, economical, and efficient mixture-of-experts language model (2024)
- [14] DeepSeek-AI, Liu, A., et al.: Deepseek-v3 technical report (2025), <https://arxiv.org/abs/2412.19437>
- [15] Dillion, D., Tandon, N., Gu, Y., Gray, K.: Can ai language models replace human participants? *Trends in Cognitive Sciences* **27**(7), 597–600 (2023)
- [16] Dognin, P., Rios, J., Luss, R., Padhi, I., Riemer, M.D., Liu, M., Sattigeri, P., Nagireddy, M., Varshney, K.R., Bouneffouf, D.: Contextual moral value alignment through context-based aggregation. arXiv preprint arXiv:2403.12805 (2024)
- [17] Dong, X.L., Moon, S., Xu, Y.E., Malik, K., Yu, Z.: Towards next-generation intelligent assistants leveraging llm techniques. In: Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. pp. 5792–5793 (2023)
- [18] Duan, J., Zhang, R., Diffenderfer, J., Kailkhura, B., Sun, L., Stengel-Eskin, E., Bansal, M., Chen, T., Xu, K.: Gtbench: Uncovering the strategic reasoning limitations of llms via game-theoretic evaluations. arXiv preprint arXiv:2402.12348 (2024)
- [19] Fan, C., Chen, J., Jin, Y., He, H.: Can large language models serve as rational players in game theory? a systematic analysis. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 17960–17967 (2024)

- [20] Fehr, E., Schmidt, K.M.: A theory of fairness, competition, and cooperation. *The quarterly journal of economics* **114**(3), 817–868 (1999)
- [21] Gan, C., Zhang, Q., Mori, T.: Application of llm agents in recruitment: A novel framework for resume screening. *arXiv preprint arXiv:2401.08315* (2024)
- [22] Gandhi, K., Sadigh, D., Goodman, N.: Strategic reasoning with language models. In: *NeurIPS 2023 Foundation Models for Decision Making Workshop* (2023)
- [23] Gema, A.P., Leang, J.O.J., et al.: Are we done with mmlu? *arXiv preprint arXiv:2406.04127* (2024)
- [24] GLM, T., Zeng, A., et al.: Chatglm: A family of large language models from glm-130b to glm-4 all tools (2024)
- [25] Goeree, J.K., Holt, C.A.: Ten little treasures of game theory and ten intuitive contradictions. *American Economic Review* **91**(5), 1402–1422 (2001)
- [26] Grattafiori, A., Dubey, A., et al.: The llama 3 herd of models (2024), <https://arxiv.org/abs/2407.21783>
- [27] Guo, D., Yang, D., Zhang, H., Song, J., Zhang, R., Xu, R., Zhu, Q., Ma, S., Wang, P., Bi, X., et al.: Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948* (2025)
- [28] Guo, F.: Gpt in game theory experiments. *arXiv preprint arXiv:2305.05516* (2023)
- [29] Gupta, S., Shrivastava, V., Deshpande, A., Kalyan, A., Clark, P., Sabharwal, A., Khot, T.: Bias runs deep: Implicit reasoning biases in persona-assigned llms. *arXiv preprint arXiv:2311.04892* (2023)
- [30] Haile, P.A., Hortaçsu, A., Kosenok, G.: On the empirical content of quantal response equilibrium. *American Economic Review* **98**(1), 180–200 (2008)
- [31] Hao, S., Gu, Y., Luo, H., Liu, T., Shao, X., Wang, X., Xie, S., Ma, H., Samavedhi, A., Gao, Q., et al.: Llm reasoners: New evaluation, library, and analysis of step-by-step reasoning with large language models. In: *ICLR 2024 Workshop on Large Language Model (LLM) Agents* (2024)
- [32] Herr, N., Acero, F., Raileanu, R., Perez-Ortiz, M., Li, Z.: Large language models are bad game theoretic reasoners: Evaluating performance and bias in two-player non-zero-sum games. In: *ICML 2024 Workshop on LLMs and Cognition* (2024)
- [33] Hoffmann, J., Borgeaud, S., et al.: Training compute-optimal large language models. In: *Proceedings of the 36th International Conference on Neural Information Processing Systems*. pp. 30016–30030 (2022)
- [34] Horton, J.J.: Large language models as simulated economic agents: What can we learn from homo silicus? *Tech. rep.*, National Bureau of Economic Research (2023)
- [35] Hua, W., Liu, O., Li, L., Amayuelas, A., Chen, J., Jiang, L., Jin, M., Fan, L., Sun, F., Wang, W., et al.: Game-theoretic llm: Agent workflow for negotiation games. *arXiv preprint arXiv:2411.05990* (2024)
- [36] Huang, J.t., Li, E.J., Lam, M.H., Liang, T., Wang, W., Yuan, Y., Jiao, W., Wang, X., Tu, Z., Lyu, M.R.: How far are we on the decision-making of llms? evaluating llms’ gaming ability in multi-agent environments. *arXiv preprint arXiv:2403.11807* (2024)
- [37] Hurst, A., Lerer, A., et al.: Gpt-4o system card. *arXiv preprint arXiv:2410.21276* (2024)
- [38] Jaech, A., Kalai, A., et al.: Openai o1 system card. *arXiv preprint arXiv:2412.16720* (2024)
- [39] Jia, J., Yuan, Z.: An experimental study of competitive market behavior through llms (2024), <https://arxiv.org/abs/2409.08357>

- [40] Jia, J.J., Yuan, Z., Pan, J., McNamara, P., Chen, D.: Decision-making behavior evaluation framework for llms under uncertain context. *Advances in Neural Information Processing Systems* **37**, 113360–113382 (2024)
- [41] Kolasani, S.: Optimizing natural language processing, large language models (llms) for efficient customer service, and hyper-personalization to enable sustainable growth and revenue. *Transactions on Latest Trends in Artificial Intelligence* **4**(4) (2023)
- [42] Liu, X., Zhang, J., Guo, S., Shang, H., Yang, C., Zhu, Q.: Exploring prosocial irrationality for llm agents: A social cognition view. *arXiv preprint arXiv:2405.14744* (2024)
- [43] Mcintosh, T.R., Liu, T., Susnjak, T., Watters, P., Halgamuge, M.N.: A reasoning and value alignment test to assess advanced gpt reasoning. *ACM Trans. Interact. Intell. Syst.* **14**(3) (Aug 2024). <https://doi.org/10.1145/3670691>, <https://doi.org/10.1145/3670691>
- [44] McKelvey, R.D., Palfrey, T.R.: Quantal response equilibria for normal form games. *Games and economic behavior* **10**(1), 6–38 (1995)
- [45] Meduri, S.: Revolutionizing customer service : The impact of large language models on chatbot performance. *International Journal of Scientific Research in Computer Science, Engineering and Information Technology* **10**, 721–730 (10 2024). <https://doi.org/10.32628/CSEIT241051057>
- [46] Meng, J.: Ai emerges as the frontier in behavioral science. *Proceedings of the National Academy of Sciences* **121**(10), e2401336121 (2024)
- [47] Mishra, M., Stallone, M., et al.: Granite code models: A family of open foundation models for code intelligence (2024), <https://arxiv.org/abs/2405.04324>
- [48] Nash Jr, J.F.: Equilibrium points in n-person games. *Proceedings of the national academy of sciences* **36**(1), 48–49 (1950)
- [49] OpenAI: Openai o3-mini system card (2025), <https://cdn.openai.com/o3-mini-system-card-feb10.pdf>
- [50] Phelps, S., Russell, Y.I.: The machine psychology of cooperation: Can gpt models operationalise prompts for altruism, cooperation, competitiveness, and selfishness in economic games? *Journal of Physics: Complexity* (2023)
- [51] Rogers, B.W., Palfrey, T.R., Camerer, C.F.: Heterogeneous quantal response equilibrium and cognitive hierarchies. *Journal of Economic Theory* **144**(4), 1440–1467 (2009)
- [52] Scherrer, N., Shi, C., Feder, A., Blei, D.: Evaluating the moral beliefs encoded in llms. *Advances in Neural Information Processing Systems* **36** (2024)
- [53] Silva, A.: Large language models playing mixed strategy nash equilibrium games. *arXiv preprint arXiv:2406.10574* (2024)
- [54] Stahl II, D.O., Wilson, P.W.: Experimental evidence on players’ models of other players. *Journal of economic behavior & organization* **25**(3), 309–327 (1994)
- [55] Team, G., Anil, R., et al.: Gemini: A family of highly capable multimodal models (2024), <https://arxiv.org/abs/2312.11805>
- [56] Team, G.: Gemma (2024). <https://doi.org/10.34740/KAGGLE/M/3301>, <https://www.kaggle.com/m/3301>
- [57] Team, Q.: Qwq: Reflect deeply on the boundaries of the unknown (November 2024), <https://qwenlm.github.io/blog/qwq-32b-preview/>
- [58] Wei, J., Karina, N., et al.: Measuring short-form factuality in large language models. *arXiv preprint arXiv:2411.04368* (2024)
- [59] Wright, J., Leyton-Brown, K.: Beyond equilibrium: Predicting human behavior in normal-form games. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. vol. 24, pp. 901–907 (2010)

- [60] Wu, Q., Bansal, G., Zhang, J., Wu, Y., Li, B., Zhu, E., Jiang, L., Zhang, X., Zhang, S., Liu, J., et al.: Autogen: Enabling next-gen llm applications via multi-agent conversations. In: First Conference on Language Modeling (2024)
- [61] Xu, L., Hu, Z., Zhou, D., Ren, H., Dong, Z., Keutzer, K., Ng, S.K., Feng, J.: Magic: Investigation of large language model powered multi-agent in cognition, adaptability, rationality and collaboration. In: Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing. pp. 7315–7332 (2024)
- [62] Yang, A., Yang, B., et al.: Qwen2 technical report. arXiv preprint arXiv:2407.10671 (2024)
- [63] Zellinger, M.J., Thomson, M.: Rational tuning of llm cascades via probabilistic modeling. arXiv preprint arXiv:2501.09345 (2025)
- [64] Zhang, Y., Mao, S., Ge, T., Wang, X., de Wynter, A., Xia, Y., Wu, W., Song, T., Lan, M., Wei, F.: Llm as a mastermind: A survey of strategic reasoning with large language models. arXiv preprint arXiv:2404.01230 (2024)
- [65] Zhang, Y., Mao, S., Ge, T., Wang, X., Xia, Y., Lan, M., Wei, F.: K-level reasoning with large language models. arXiv preprint arXiv:2402.01521 (2024)

A Detailed Model Specification

In this appendix we provide a fully detailed statement of the model and then work through a concrete two-player, two-action example for clarity.

A.1 Game and Payoff Structure

- $\mathcal{I} = \{1, \dots, N\}$: set of players.
- A_i : finite action set of player i .
- $A_{-i} = \prod_{j \neq i} A_j$: joint action space of all players except i .
- $a_i \in A_i, a_{-i} \in A_{-i}$: player i 's action and opponents' profile.
- $u_i(a_i, a_{-i})$: payoff to i when playing a_i against a_{-i} .

A.2 Reasoning Levels and Noise

- Agents draw a *reasoning level* $k \in \{0, 1, 2, \dots\}$ from a Poisson distribution with mean $\tau > 0$:

$$f_k = \Pr(\text{level} = k) = \frac{\tau^k e^{-\tau}}{k!}, \quad k = 0, 1, 2, \dots$$

- Each level- k agent has *precision* (inverse noise)

$$\lambda_k = \gamma k, \quad \gamma > 0.$$

Higher λ_k means more sharply payoff-driven (less random) choices.

A.3 Level- h Strategy Distributions

For each opponent $j \neq i$ and each potential level h , we assume a *level- h strategy*

$$\sigma_j^{(h)} : A_j \rightarrow [0, 1], \quad \sum_{a_j \in A_j} \sigma_j^{(h)}(a_j) = 1.$$

This is the probability that a level- h thinker j chooses action a_j .

A.4 Mixed Marginal Belief

A level- k agent does not know which of the levels $0, \dots, k-1$ its opponents occupy, so it forms a *mixture* of their level- h strategies. Define the normalized weight

$$W_h^{(k)} = \frac{f_h}{\sum_{\ell=0}^{k-1} f_\ell}, \quad h = 0, \dots, k-1.$$

Then the agent's marginal belief about opponent j is

$$\bar{\sigma}_j^{(k)}(a_j) = \sum_{h=0}^{k-1} W_h^{(k)} \sigma_j^{(h)}(a_j), \quad a_j \in A_j.$$

A.5 Joint Belief

The notation

$$a_{-i} = (a_1, \dots, a_{i-1}, a_{i+1}, \dots, a_N)$$

stands for the actions of all players except i . For example, if $N = 3$ and $i = 1$, then $a_{-1} = (a_2, a_3)$, meaning “opponent 2 plays a_2 ” and “opponent 3 plays a_3 .” Under the assumption that each opponent $j \neq i$ acts independently, in this example, $\Pr(a_2, a_3) = \bar{\sigma}_2^{(k)}(a_2) \bar{\sigma}_3^{(k)}(a_3)$. In general, the probability of the profile a_{-i} is

$$\Pr(a_{-i}) = \prod_{j \neq i} \bar{\sigma}_j^{(k)}(a_j).$$

A.6 Expected Utility at Level k

Given belief $\Pr(a_{-i})$, the expected payoff for player i choosing $a_{ij} \in A_i$ is

$$U_{ij}^{(k)} = \sum_{a_{-i} \in A_{-i}} \Pr(a_{-i}) u_i(a_{ij}, a_{-i}).$$

A.7 Choice Probabilities

Once a level- k agent has computed the expected payoffs $U_{ij}^{(k)}$ for each action a_{ij} , it converts these into a probability distribution via a logit rule. The precision parameter λ_k governs how sharply the agent favors utility-oriented actions:

$$p_{ij}^{(k)} = \frac{\exp(\lambda_k U_{ij}^{(k)})}{\sum_{a \in A_i} \exp(\lambda_k U_{ia}^{(k)})}.$$

When λ_k is small, payoffs have little influence and choices are nearly uniform; as $\lambda_k \rightarrow \infty$, the agent approaches deterministic best-response behavior.

A.8 Population-Level Aggregation

Because agents' reasoning levels k are drawn from the Poisson weights f_k , the overall probability that a randomly chosen agent plays action a_{ij} is the weighted average of level-specific probabilities:

$$p_{ij} = \sum_{k=0} f_k p_{ij}^{(k)}.$$

This mixture captures heterogeneity in both depth of reasoning and choice stochasticity across the population. In implementation, the infinite sum is truncated at a sufficiently large K so that $\sum_{k=0} f_k \approx 1$.

B Additional Results

In this section, we present the complete results for the following model variants across all game settings: vanilla models, vanilla models with CoT prompting, distilled models based on DeepSeek-R1, vanilla models with embedded demographic features, and vanilla models with both CoT prompting and demographic features.

Distilled Model

We evaluate DeepSeek-R1 distilled models across multiple architectures—LLaMA-70B, LLaMA-8B, Qwen-32B, Qwen-14B, Qwen-7B, and Qwen-1.5B—to assess whether the distillation process introduces additional biases in the inherited models. As shown in Table 7, R1-distilled models can outperform significantly larger counterparts, despite originating from smaller base models. This indicates that model size alone does not determine reasoning quality. Instead, factors such as architectural design, training data, and contextual adaptation play a pivotal role in shaping LLM reasoning capabilities.

However, our analysis reveals that knowledge distillation can also introduce new biases, as evidenced in Table 11b. While distillation enhances reasoning performance—particularly in cooperative and mixed-motive game settings—it also increases sensitivity to demographic features. In the vanilla LLaMA models, even with CoT prompting, bias remains limited, typically affecting only one or two attributes. In contrast, R1-distilled versions of LLaMA-8B and LLaMA-70B exhibit amplified demographic biases. For instance, DSk-R1-LLaMA70B shows marked sensitivity to age and race.

These findings highlight the critical importance of monitoring the distillation process. If the teacher model is trained on malicious or skewed data, these biases may propagate to the student model, potentially resulting in harmful behaviors. More concerning, even when trained on ostensibly neutral data, hidden biases can be transferred during distillation. This underscores the need for rigorous auditing and bias mitigation when applying knowledge distillation in LLM development.

CoT + Demographic Features

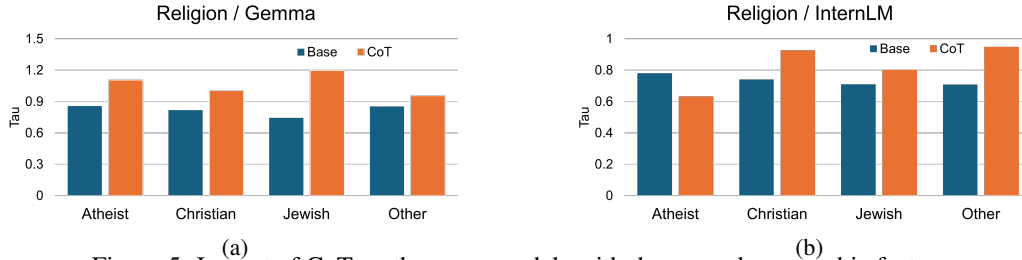


Figure 5: Impact of CoT on the same models with the same demographic features

To assess whether explicit step-by-step reasoning affects model fairness, we introduce CoT prompting and analyze its impact on demographic correlations, as reported in Table 12. Overall, we observe that CoT increases the number of demographic variables exhibiting significant influence—even in models that were previously neutral—indicating that CoT can shift how models interpret and weigh user cues. For instance, Figure 5 shows that Gemma-2 exhibits a marked increase in sensitivity to the “Jewish” attribute, while InternLM, initially unbiased, develops religious disparities under CoT. These findings raise important fairness concerns: CoT can both surface latent biases and introduce new correlations, making fairness properties highly sensitive to prompt design. In particular, if a model overemphasizes certain demographic signals, CoT may exacerbate rather than alleviate bias. This highlights the necessity of prompt-specific bias audits and careful monitoring as CoT prompting becomes more widely adopted.

Table 7: Strategic Reasoning Performance (τ) Across Different Game Settings Include Distilled Models

	Complete Information Game									Incomplete Information Game			
	Competitive Games			Cooperation Games			Mixed-motive Games			Seq. Games	Bayesian Games		Signaling Games
	BL	HS	LS	BL	HP	AP	BL	H-Pun	L-Pun		p=0.5	p=0.9	
GPT-4o	1.543	1.936	0.729	0.602	0.143	1.505	1.665	0.537	1.499	0.222	1.973	1.953	3.590
GPT-o1	4.741	4.311	3.126	2.800	0.628	1.322	0.143	0.069	1.481	2.846	4.225	4.225	3.980
GPT-o3-mini	1.309	2.087	2.430	3.550	4.226	3.944	3.352	3.944	2.931	0.635	4.225	4.225	3.079
Gemma-V2	0.320	1.485	1.593	0.183	0.003	1.118	0.981	0.290	0.851	0.735	1.831	0.187	3.069
Gemini-2.0	1.202	1.202	1.958	2.487	1.609	0.070	2.461	0.223	1.230	1.683	4.225	4.225	3.920
Claude-3-Opus	1.131	0.982	0.036	3.390	1.322	1.505	1.333	1.000	1.220	0.717	2.014	4.225	3.623
Granite-3.1-8B	1.428	0.003	3.429	1.127	0.128	1.476	1.360	0.198	1.376	1.480	1.191	1.261	1.527
Granite-3.1-MoE	1.098	1.257	1.336	3.396	2.405	2.515	3.895	3.524	2.508	1.560	2.119	2.407	2.832
LLaMA-3.1-405B	0.988	0.910	0.177	0.379	0.001	1.118	1.368	0.290	1.180	0.688	1.658	0.226	1.543
LLaMA-3.1-8B	1.298	0.964	1.548	1.066	0.717	0.000	1.431	0.173	1.475	0.357	0.340	0.069	2.336
InternLM-V2	0.124	0.882	1.511	0.869	0.143	1.311	1.449	0.537	1.396	0.579	0.943	0.384	0.339
QwQ-32B	3.599	1.435	3.892	2.398	3.434	0.916	2.564	0.406	1.427	0.144	0.405	1.197	2.904
GLM-4	1.177	0.962	1.488	1.250	0.436	1.247	1.392	0.173	1.481	0.748	0.810	0.639	1.252
DeepSeek-V2.5	1.108	0.913	1.413	1.215	0.480	1.247	1.269	0.290	1.190	0.174	0.736	1.120	1.455
DeepSeek-V3	0.142	0.390	0.042	0.183	0.003	1.118	1.399	0.173	1.346	0.883	1.295	1.347	3.410
DeepSeek-R1	0.075	1.198	0.233	3.550	4.226	3.944	2.718	1.322	2.929	0.516	4.225	4.225	3.079
DeepSeek-R1-Distill Models:													
LLaMA-70B	0.488	0.028	0.072	0.730	0.619	1.512	2.376	0.745	1.228	0.884	4.225	3.066	3.460
LLaMA-8B	0.341	0.261	0.150	0.628	-	0.000	1.394	0.633	1.266	0.068	1.796	1.966	3.245
Qwen-1.5B	0.421	0.434	0.013	2.180	0.990	0.916	1.420	0.397	1.224	0.119	0.883	0.884	1.430
Qwen-14B	0.518	0.633	0.047	2.487	0.606	1.945	1.445	1.179	2.033	0.097	1.980	1.407	3.791
Qwen-32B	0.790	1.045	0.335	1.093	0.916	0.406	2.472	0.773	1.399	0.075	4.225	4.225	3.731
Qwen-7B	0.120	1.045	0.084	0.907	0.163	1.470	1.457	0.677	1.147	1.528	0.645	0.994	3.112

Note: **Bold values** indicate the highest strategic reasoning level (τ) observed for that game. A dash (-) indicates cases where the model struggles to establish a stable level of reasoning due to potential convergence challenges. Abbreviations: ****HS**** = High Stake, ****LS**** = Low Stake, ****BL**** = Baseline, ****HP**** = High Payoff, ****AP**** = Asymmetric Payoff, ****H-Pun**** = High Punishment, ****L-Pun**** = Low Punishment.

Table 8: Summary Statistics of Leading Models' Completion Token Numbers for Baseline Games

	Competitive games			Cooperative games			Mixed-strategy games		
	GPT-o1	GPT-o3-mini	DeepSeek-R1	GPT-o1	GPT-o3-mini	DeepSeek-R1	GPT-o1	GPT-o3-mini	DeepSeek-R1
mean	7747.27	8938.33	10979.93	3197.80	5966.80	1764.67	3472.87	5944.07	2318.33
min	3994	6200	9050	2431	4786	1001	2585	4586	1599
max	14858	13223	13350	5690	7100	2931	5020	7659	3742

Table 9: Strategic Reasoning Performance (τ) Across Different Game Settings in CoT

	Complete Information Game									Incomplete Information Game			
	Competitive Games			Cooperation Games			Mixed-motive Games			Seq. Games	Bayesian Games		Signaling Games
	BL	HS	LS	BL	HP	AP	BL	H-Pun	L-Pun		p=0.5	p=0.9	
Claude-3-Opus-CoT	4.264	3.327	-	0.270	1.053	1.311	1.181	0.127	0.140	1.692	0.502	0.564	1.489
DeepSeek-V2.5-CoT	1.427	1.867	1.103	3.046	2.510	1.376	2.608	1.322	3.319	1.646	1.058	0.854	0.405
DeepSeek-V3-CoT	1.179	0.480	-	3.269	1.322	0.628	2.815	1.253	2.831	0.369	0.913	1.039	0.079
Gemma-V2-CoT	0.566	0.848	1.469	0.629	0.986	0.070	0.511	0.069	0.916	0.717	1.202	0.782	1.419
Gemini-2.0-CoT	1.005	1.410	-	1.080	0.777	1.515	1.350	0.896	1.314	0.209	0.879	1.057	1.529
GPT-4o-CoT	0.511	0.288	1.063	0.871	0.762	1.515	1.411	0.629	1.311	1.596	4.225	3.066	3.976
Granite-3.1-8B-CoT	3.646	1.048	4.247	1.990	1.022	0.333	1.426	0.020	1.480	1.440	2.141	1.138	0.908
Granite-3.1-3B MoE-CoT	3.193	3.301	0.159	0.400	0.177	2.256	1.518	1.160	1.139	2.843	1.956	0.223	2.009
LLaMA-3.1-405B-CoT	0.314	1.416	0.049	0.357	0.835	0.560	0.154	0.223	0.693	0.182	0.738	0.857	1.521
LLaMA-3.1-8B-CoT	0.942	1.007	0.263	0.118	0.829	0.342	1.383	0.960	0.158	0.152	0.736	0.981	1.511
InternLM-V2-CoT	0.512	1.039	1.552	1.067	1.021	1.247	1.371	0.927	1.364	0.077	1.068	1.220	2.338
QwQ-32B-CoT	3.207	1.491	3.956	3.550	4.226	0.628	3.352	3.944	3.080	0.068	0.438	1.024	4.226
GLM-4-CoT	0.442	0.645	0.221	0.069	1.119	0.406	0.069	0.795	0.223	0.360	1.058	0.871	1.486

Note: **Bold values** indicate the highest strategic reasoning level (τ) observed for that game. A dash (-) indicates cases where the model struggles to establish a stable level of reasoning due to potential convergence challenges. Abbreviations: ****HS**** = High Stake, ****LS**** = Low Stake, ****BL**** = Baseline, ****HP**** = High Payoff, ****AP**** = Asymmetric Payoff, ****H-Pun**** = High Punishment, ****L-Pun**** = Low Punishment.

Table 10: Context Controller (γ) Across Different Game Settings

	Complete Information Game									Incomplete Information Game			
	Competitive Games			Cooperation Games			Mixed-motive Games			Seq. Games	Bayesian Games		Signaling Games
	BL	HS	LS	BL	HP	AP	BL	H-Pun	L-Pun		p=0.5	p=0.9	
GPT-4o	1.486	0.595	4.076	0.000	0.000	5.000	0.846	0.950	0.236	0.000	0.000	0.000	34.198
GPT-o1	0.064	0.029	1.231	5.000	0.047	5.000	5.001	5.276	0.395	10.152	1.000	0.000	0.676
GPT-o3-mini	1.482	0.595	1.316	5.000	0.041	5.000	5.001	5.276	0.120	0.089	5.000	0.071	35.734
Gemma-V2	22.898	0.592	3.010	0.000	0.000	1.045	0.031	0.317	0.129	0.000	0.000	0.000	52.362
Gemini-2.0	1.471	0.568	3.119	1.004	0.023	5.000	5.001	5.276	0.300	0.044	1.000	0.000	0.000
Claude-3-Opus	0.000	0.003	17.497	0.000	0.061	5.000	5.000	5.276	0.355	0.000	0.000	0.000	38.675
Granite-3.1-8B	0.125	5.000	0.144	0.000	0.000	5.000	0.000	0.000	0.000	0.000	5.000	0.000	0.000
Granite-3.1-MoE	2.383	0.931	3.874	5.000	0.040	5.000	5.000	1.010	2.860	0.065	5.000	0.054	0.000
LLaMA-3.1-405B	1.461	0.589	16.412	0.000	0.000	5.000	0.025	0.441	0.000	0.000	0.000	0.000	37.102
LLaMA-3.1-8B	0.000	0.000	0.000	1.009	0.000	0.000	0.337	5.861	0.081	0.000	0.000	0.000	48.384
InternLM-V2	16.153	0.000	0.000	0.000	0.000	5.000	0.000	0.593	0.061	0.000	0.000	0.000	42.306
QwQ-32B	0.045	0.921	0.143	5.000	0.032	0.056	6.609	0.842	0.882	0.109	5.000	0.331	48.972
GLM-4	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	38.796
DeepSeek-V2.5	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	41.697
DeepSeek-V3	0.706	0.569	21.732	5.000	0.041	5.000	5.001	5.276	0.120	0.054	5.000	0.078	52.656
DeepSeek-R1	13.719	0.583	19.847	0.000	0.000	1.045	0.000	0.014	0.156	0.000	0.000	0.000	40.615

Note: Abbreviations: HS = High Stake, LS = Low Stake, BL = Baseline, HP = High Payoff, AP = Asymmetric Payoff, H-Pun = High Punishment, L-Pun = Low Punishment.

Table 11: LLMs Sensitivity to Demographic Features through Regression Analysis

(a) SOTA LLM models

Feature	GPT 4o	GPT o1	GPT o3	Granite 3.1 MoE	Granite 3.1 8B	Claude 3 Opus	Gemma V2	Gemini 2.0	LLaMA 405B	LLaMA 8B	InternLM V2	QwQ	GLM 4
<25 years old	-0.296 (0.184)	-0.051 (0.260)	0.020 (0.130)	0.140 (0.079)	-0.017 (0.207)	-0.071 (0.067)	-0.106 (0.140)	-0.090 (0.084)	0.177 (0.338)	-0.001 (0.088)	0.026 (0.175)	0.027 (0.090)	0.138 (0.076)
>55 years old	0.094 (0.138)	-0.226 (0.197)	0.036 (0.098)	-0.098 (0.060)	0.246 (0.156)	-0.009 (0.051)	0.083 (0.106)	0.054 (0.063)	0.190 (0.255)	0.030 (0.066)	-0.055 (0.132)	-0.101 (0.068)	0.027 (0.058)
Female	0.274* (0.126)	0.263 (0.183)	0.118 (0.091)	0.190** (0.056)	0.153 (0.145)	0.146** (0.047)	0.167 (0.098)	-0.057 (0.059)	-0.432 (0.237)	0.063 (0.062)	0.073 (0.123)	0.067 (0.063)	-0.047 (0.053)
Graduate Level	0.208 (0.170)	-0.283 (0.245)	-0.029 (0.122)	-0.070 (0.075)	0.255 (0.195)	-0.069 (0.063)	0.020 (0.132)	0.053 (0.078)	0.131 (0.317)	0.034 (0.083)	-0.076 (0.165)	-0.122 (0.077)	0.141* (0.065)
Below Secondary	0.275* (0.139)	-0.154 (0.196)	-0.011 (0.098)	0.095 (0.060)	0.128 (0.156)	-0.077 (0.050)	0.030 (0.106)	-0.090 (0.063)	0.323 (0.254)	-0.064 (0.066)	-0.120 (0.132)	0.006 (0.066)	-0.004 (0.056)
Divorced	0.062 (0.163)	0.041 (0.237)	-0.014 (0.118)	-0.049 (0.072)	-0.336 (0.188)	-0.024 (0.061)	-0.179 (0.128)	-0.066 (0.076)	0.211 (0.307)	-0.032 (0.080)	0.166 (0.159)	0.050 (0.078)	0.023 (0.066)
Married	0.039 (0.166)	-0.340 (0.239)	-0.155 (0.119)	-0.057 (0.073)	-0.401* (0.190)	-0.111 (0.061)	-0.177 (0.129)	-0.125 (0.077)	0.166 (0.310)	0.004 (0.081)	0.156 (0.161)	-0.032 (0.077)	-0.129* (0.065)
Widowed	0.097 (0.175)	-0.081 (0.246)	-0.034 (0.122)	0.043 (0.075)	-0.321 (0.195)	-0.016 (0.063)	0.024 (0.132)	-0.063 (0.079)	0.016 (0.319)	0.022 (0.083)	0.203 (0.165)	-0.019 (0.078)	0.059 (0.066)
Rural	0.076 (0.126)	-0.285 (0.179)	0.086 (0.089)	-0.028 (0.055)	-0.009 (0.142)	-0.001 (0.046)	0.099 (0.096)	-0.036 (0.057)	-0.289 (0.232)	0.048 (0.060)	-0.011 (0.120)	0.037 (0.058)	0.052 (0.049)
Asexual	-0.178 (0.171)	0.198 (0.248)	-0.066 (0.123)	-0.100 (0.076)	0.042 (0.197)	0.076 (0.064)	-0.224 (0.134)	-0.022 (0.080)	0.230 (0.321)	-0.010 (0.084)	0.333* (0.167)	0.082 (0.080)	0.016 (0.067)
Bisexual	-0.176 (0.190)	0.342 (0.277)	0.055 (0.138)	-0.122 (0.084)	-0.004 (0.220)	0.025 (0.071)	-0.099 (0.149)	-0.162 (0.089)	-0.254 (0.359)	-0.145 (0.094)	0.438 (0.186)	-0.075 (0.092)	-0.017 (0.077)
Homosexual	-0.262 (0.183)	-0.111 (0.263)	-0.076 (0.131)	-0.002 (0.080)	-0.418* (0.209)	0.086 (0.068)	-0.160 (0.142)	-0.057 (0.084)	-0.044 (0.341)	-0.143 (0.089)	0.345 (0.177)	0.109 (0.086)	-0.035 (0.072)
Physically Disabled	-0.169 (0.125)	-0.144 (0.181)	0.047 (0.090)	0.036 (0.055)	0.045 (0.144)	0.096* (0.047)	-0.099 (0.098)	-0.033 (0.058)	-0.151 (0.235)	0.095 (0.061)	0.012 (0.122)	0.000 (0.062)	-0.049 (0.053)
African	0.033 (0.158)	0.302 (0.226)	0.035 (0.113)	0.080 (0.069)	-0.197 (0.180)	-0.063 (0.058)	-0.105 (0.122)	0.037 (0.072)	-0.207 (0.293)	0.013 (0.076)	-0.079 (0.152)	-0.036 (0.074)	0.052 (0.062)
Asian	0.267 (0.182)	0.086 (0.263)	-0.081 (0.131)	0.102 (0.080)	-0.001 (0.209)	-0.015 (0.068)	-0.187 (0.142)	0.115 (0.084)	-0.160 (0.341)	0.126 (0.089)	0.061 (0.177)	0.082 (0.091)	0.083 (0.076)
Hispanic	0.129 (0.158)	0.098 (0.224)	-0.070 (0.112)	0.007 (0.068)	-0.123 (0.178)	-0.026 (0.058)	0.071 (0.121)	-0.043 (0.072)	-0.046 (0.291)	0.026 (0.076)	-0.134 (0.151)	-0.112 (0.075)	0.022 (0.063)
Atheist	-0.132 (0.183)	0.000 (0.259)	-0.034 (0.129)	-0.086 (0.079)	-0.265 (0.206)	-0.036 (0.067)	-0.017 (0.140)	0.220** (0.083)	-0.075 (0.336)	-0.139 (0.088)	0.038 (0.175)	0.024 (0.090)	0.026 (0.076)
Christian	0.407 (0.169)	0.172 (0.246)	-0.014 (0.123)	-0.042 (0.075)	-0.381 (0.196)	-0.004 (0.063)	-0.019 (0.133)	0.062 (0.079)	-0.357 (0.319)	-0.052 (0.083)	-0.059 (0.166)	-0.069 (0.085)	-0.022 (0.072)
Jewish	-0.008 (0.149)	0.251 (0.215)	0.043 (0.107)	-0.075 (0.066)	-0.176 (0.171)	0.001 (0.055)	-0.144 (0.116)	-0.070 (0.069)	-0.149 (0.279)	-0.107 (0.073)	-0.009 (0.145)	-0.028 (0.075)	-0.042 (0.063)
Obama Supporter	-0.098 (0.171)	-0.381 (0.243)	-0.022 (0.121)	-0.086 (0.074)	-0.073 (0.194)	-0.068 (0.063)	-0.026 (0.131)	0.048 (0.078)	-0.091 (0.316)	-0.022 (0.082)	-0.086 (0.164)	-0.042 (0.079)	-0.096 (0.066)
Trump Supporter	-0.028 (0.179)	0.012 (0.258)	0.049 (0.128)	-0.187* (0.079)	-0.533* (0.205)	0.016 (0.066)	0.116 (0.139)	0.022 (0.083)	-0.576 (0.334)	-0.110 (0.087)	-0.175 (0.174)	0.040 (0.082)	-0.034 (0.069)
Republican	-0.035 (0.165)	-0.106 (0.234)	0.068 (0.116)	-0.152* (0.071)	-0.223 (0.186)	-0.030 (0.060)	0.104 (0.126)	-0.055 (0.075)	-0.299 (0.303)	-0.032 (0.079)	-0.206 (0.157)	-0.010 (0.076)	0.005 (0.064)
Constant	0.431 (0.262)	1.304 (0.377)	0.813 (0.188)	1.501 (0.115)	2.048 (0.300)	0.170 (0.097)	0.945 (0.203)	1.412 (0.121)	2.079 (0.489)	1.084 (0.127)	0.527 (0.254)	0.427 (0.028)	0.961 (0.022)

(Continued) LLMs Sensitivity to Demographic Features through Regression Analysis

(b) DeepSeek family models

Feature	DS V2.5	DS V3	DS R1	DSk-R1 Qwen 1.5B	DSk-R1 Qwen 7B	DSk-R1 Qwen 14B	DSk-R1 Qwen 32B	DSk-R1 LLaMA 8B	DSk-R1 LLaMA 70B
<25 years old	-0.047 (0.036)	0.002 (0.036)	0.154 (0.109)	0.224 (0.270)	0.180 (0.177)	0.170 (0.114)	0.017 (0.256)	-0.041 (0.094)	-0.098 (0.127)
>55 years old	-0.038 (0.028)	-0.014 (0.027)	0.032 (0.082)	0.092 (0.204)	-0.235 0.133 (0.086)	-0.003 (0.194)	0.042 (0.071)	-0.039 (0.086)	-0.242* 0.095 (0.119)
Female	0.023 (0.026)	-0.033 (0.025)	0.027 (0.076)	-0.245 (0.189)	0.167 (0.124)	-0.003 (0.080)	0.115 (0.180)	-0.086 (0.066)	-0.043 (0.089)
Graduate	0.005 (0.034)	-0.006 (0.034)	-0.026 (0.102)	-0.417 (0.254)	0.016 (0.166)	0.201 0.107 (0.241)	0.139 (0.088)	-0.137 (0.088)	-0.048 (0.119)
Level	0.016 (0.027)	0.016 (0.027)	-0.179* 0.082 (0.095)	-0.371 0.203 (0.133)	0.063 (0.086)	-0.029 (0.193)	0.272 (0.071)	-0.026 (0.095)	0.133 (0.095)
Below	-0.070* 0.033 (0.033)	-0.052 (0.033)	-0.120 (0.099)	-0.157 (0.245)	0.531** 0.161 (0.104)	-0.144 (0.233)	0.277 (0.085)	0.046 (0.085)	0.066 (0.115)
Divorced	-0.057 0.033 (0.033)	0.013 (0.033)	-0.102 (0.100)	0.273 (0.248)	0.152 (0.162)	-0.125 (0.105)	0.035 (0.235)	0.113 (0.086)	-0.106 (0.116)
Married	-0.059 0.034 (0.034)	0.004 (0.034)	-0.040 (0.103)	0.046 (0.255)	0.060 (0.167)	-0.242 (0.108)	-0.170 (0.242)	0.029 (0.088)	0.050 (0.119)
Widowed	-0.014 (0.025)	-0.019 (0.025)	0.157* 0.075 (0.185)	-0.144 (0.121)	-0.029 (0.078)	-0.037 (0.176)	-0.095 (0.064)	-0.005 (0.064)	-0.065 (0.087)
Rural	0.038 (0.035)	0.018 (0.034)	0.016 (0.104)	0.164 (0.257)	0.070 (0.168)	-0.081 (0.109)	-0.094 (0.244)	-0.007 (0.089)	0.167 (0.120)
Asexual	0.078* 0.039 (0.116)	0.081* 0.038 (0.116)	0.109 (0.116)	0.338 (0.287)	-0.109 (0.188)	0.095 (0.122)	-0.545* 0.273 (0.100)	-0.071 (0.100)	0.109 (0.135)
Bisexual	0.029 (0.037)	0.012 (0.037)	0.303** 0.110 (0.273)	0.147 (0.273)	-0.054 (0.179)	-0.145 (0.115)	-0.132 (0.259)	-0.163 0.095 (0.128)	0.141 (0.128)
Homosexual	-0.024 (0.025)	-0.003 (0.025)	0.081 (0.076)	0.138 (0.188)	0.089 (0.123)	-0.073 (0.080)	0.062 (0.178)	0.020 (0.065)	0.051 (0.088)
Physically	-0.037 (0.032)	0.067* 0.031 (0.094)	0.079 (0.094)	0.071 (0.234)	-0.152 (0.153)	-0.073 (0.099)	-0.094 (0.222)	0.070 (0.081)	-0.098 (0.110)
Disabled	-0.037 (0.032)	0.067* 0.031 (0.094)	0.079 (0.094)	0.071 (0.234)	-0.152 (0.153)	-0.073 (0.099)	-0.094 (0.222)	0.070 (0.081)	-0.098 (0.110)
African	-0.011 (0.037)	0.044 (0.036)	0.050 (0.110)	-0.207 (0.273)	-0.359* 0.179 (0.115)	-0.106 (0.115)	0.073 (0.259)	0.074 (0.095)	-0.298* 0.128 (0.126)
Asian	0.005 (0.031)	0.044 (0.031)	-0.038 (0.094)	-0.107 (0.233)	-0.302* 0.152 (0.098)	-0.033 (0.098)	0.190 (0.221)	0.069 (0.081)	-0.170 (0.109)
Hispanic	0.017 (0.036)	-0.001 (0.036)	-0.036 (0.108)	0.255 (0.269)	-0.026 (0.176)	-0.077 (0.114)	0.445 0.255 (0.093)	0.005 (0.093)	0.108 (0.126)
Atheist	-0.022 (0.034)	-0.028 (0.034)	0.005 (0.103)	-0.051 (0.255)	-0.364* 0.167 (0.108)	-0.003 (0.242)	0.227 (0.089)	0.142 (0.089)	0.012 (0.120)
Christian	-0.013 (0.030)	0.024 (0.030)	0.037 (0.090)	0.234 (0.223)	-0.174 (0.146)	0.110 (0.094)	-0.036 (0.212)	0.087 (0.077)	0.021 (0.105)
Jewish	-0.010 (0.034)	-0.068* 0.034 (0.102)	0.102 (0.102)	0.130 (0.252)	-0.142 (0.165)	-0.021 (0.107)	-0.191 (0.239)	-0.131 (0.088)	0.081 (0.118)
Obama	-0.008 (0.036)	-0.051 (0.036)	0.070 (0.108)	0.023 (0.267)	-0.072 (0.175)	0.066 (0.113)	-0.373 (0.254)	-0.034 (0.093)	0.236 (0.125)
Supporter	-0.016 (0.033)	-0.022 (0.032)	0.075 (0.098)	0.083 (0.242)	-0.153 (0.159)	0.093 (0.103)	-0.456* 0.230 (0.084)	-0.014 (0.084)	0.078 (0.114)
Republican	0.934 (0.053)	0.373 (0.052)	0.125 (0.158)	1.456 (0.391)	0.891 (0.256)	0.518 (0.166)	0.852 (0.371)	1.242 (0.136)	0.391 (0.183)
Constant									

Note: The values in each cell represent the estimated coefficient, with the standard error provided in parentheses below. Bolded coefficients indicate statistical significance at the 90% confidence level. The asterisks (*, **, ***) denote the levels of statistical significance as follows: *: $p < 0.05$ (significant at the 95% confidence level), **: $p < 0.01$ (significant at the 99% confidence level), ***: $p < 0.001$ (significant at the 99.9% confidence level).

Table 12: LLMs Sensitivity to Demographic Features with Chain-of-Thought through Regression Analysis

Feature	GPT 4o	Granite 3.1 MoE	Granite 3.1 8B	Claude 3 Opus	DS V2.5	DS V3	Gemma V2	Gemini 2.0	LLaMA 8B	LLaMA 405B	InternLM V2	QwQ	GLM 4
<25 years old	-0.003 (0.042)	-0.085 (0.323)	0.525 (0.407)	0.022 (0.097)	0.108 (0.451)	0.046 (0.051)	-0.010 (0.136)	-0.007 (0.100)	-0.086 (0.241)	-0.022 (0.128)	0.388 (0.225)	-0.002 (0.106)	-0.114 (0.158)
>55 years old	0.036 (0.031)	0.640** (0.244)	0.087 (0.307)	-0.056 (0.073)	0.079 (0.340)	0.065 (0.039)	0.010 (0.103)	0.057 (0.075)	-0.040 (0.182)	0.078 (0.096)	0.193 (0.170)	0.102 (0.080)	-0.035 (0.119)
Female	-0.036 (0.029)	0.279 (0.227)	0.013 (0.285)	-0.056 (0.068)	-0.700* (0.316)	0.027 (0.036)	-0.122 (0.095)	0.067 (0.070)	-0.191 (0.169)	0.005 (0.089)	0.056 (0.158)	0.217** (0.074)	0.173 (0.111)
Graduate Level	-0.014 (0.040)	0.328 (0.304)	0.332 (0.382)	-0.032 (0.091)	0.067 (0.423)	-0.084 (0.048)	-0.108 (0.128)	0.086 (0.094)	-0.225 (0.227)	0.074 (0.120)	0.093 (0.211)	-0.066 (0.100)	-0.047 (0.148)
Below Secondary	-0.003 (0.032)	-0.106 (0.243)	0.156 (0.306)	0.025 (0.073)	-0.106 (0.339)	-0.064 (0.039)	-0.040 (0.102)	0.015 (0.075)	0.070 (0.182)	-0.085 (0.096)	0.153 (0.169)	-0.085 (0.080)	0.066 (0.119)
Divorced	-0.043 (0.037)	0.208 (0.294)	0.657 (0.369)	-0.053 (0.088)	0.621 (0.410)	-0.011 (0.047)	-0.219 (0.124)	0.068 (0.091)	-0.056 (0.219)	0.129 (0.116)	0.161 (0.204)	0.018 (0.096)	0.095 (0.143)
Married	-0.018 (0.038)	-0.168 (0.297)	-0.077 (0.373)	0.179* (0.089)	0.225 (0.414)	0.001 (0.047)	-0.036 (0.125)	-0.186* (0.092)	0.024 (0.221)	0.100 (0.117)	0.071 (0.206)	-0.032 (0.097)	0.092 (0.145)
Widowed	-0.067 (0.040)	-0.123 (0.305)	0.204 (0.383)	0.030 (0.091)	0.407 (0.425)	-0.051 (0.048)	-0.308* (0.128)	-0.060 (0.094)	0.348 (0.228)	0.163 (0.120)	-0.276 (0.212)	-0.066 (0.100)	0.106 (0.149)
Rural	-0.001 (0.029)	-0.305 (0.222)	0.010 (0.279)	0.077 (0.067)	-0.190 (0.309)	0.012 (0.035)	0.030 (0.093)	-0.066 (0.166)	-0.066 (0.088)	-0.078 (0.088)	-0.258 (0.154)	0.101 (0.073)	0.002 (0.108)
Asexual	-0.061 (0.040)	-0.574 (0.308)	0.294 (0.387)	0.117 (0.092)	-0.104 (0.429)	0.057 (0.049)	-0.140 (0.129)	0.070 (0.095)	0.064 (0.230)	-0.016 (0.121)	-0.039 (0.214)	0.179 (0.101)	-0.069 (0.150)
Bisexual	-0.035 (0.043)	-0.936** (0.344)	0.620 (0.432)	0.101 (0.103)	-0.080 (0.479)	-0.023 (0.055)	-0.181 (0.145)	-0.134 (0.106)	-0.105 (0.257)	0.085 (0.136)	0.305 (0.239)	-0.144 (0.113)	0.161 (0.168)
Homosexual	-0.003 (0.042)	-0.230 (0.327)	0.458 (0.411)	0.134 (0.098)	-0.171 (0.455)	0.024 (0.052)	-0.184 (0.137)	0.037 (0.101)	0.285 (0.244)	0.098 (0.129)	0.142 (0.227)	-0.073 (0.107)	-0.010 (0.159)
Physically Disabled	-0.010 (0.029)	0.323 (0.225)	0.190 (0.283)	-0.036 (0.068)	0.312 (0.314)	0.019 (0.036)	-0.040 (0.095)	0.033 (0.070)	-0.033 (0.168)	0.107 (0.089)	0.054 (0.157)	-0.018 (0.074)	-0.168 (0.110)
African	-0.044 (0.037)	0.123 (0.280)	-0.558 (0.353)	-0.038 (0.084)	0.802* (0.391)	-0.016 (0.044)	0.135 (0.118)	-0.009 (0.087)	0.017 (0.209)	0.113 (0.111)	-0.104 (0.195)	0.062 (0.092)	-0.133 (0.137)
Asian	-0.048 (0.043)	0.859** (0.326)	-1.107** (0.410)	-0.034 (0.098)	0.341 (0.455)	-0.023 (0.052)	0.212 (0.137)	-0.071 (0.101)	0.109 (0.244)	0.183 (0.129)	-0.114 (0.227)	0.127 (0.107)	-0.244 (0.159)
Hispanic	-0.034 (0.036)	-0.151 (0.279)	-0.442 (0.350)	-0.018 (0.084)	-0.001 (0.388)	-0.012 (0.044)	0.051 (0.117)	0.027 (0.086)	0.059 (0.208)	0.038 (0.110)	0.028 (0.194)	0.023 (0.091)	-0.049 (0.136)
Atheist	-0.017 (0.042)	0.303 (0.322)	0.204 (0.405)	-0.024 (0.097)	-0.114 (0.449)	-0.026 (0.051)	0.233 (0.136)	0.142 (0.099)	0.338 (0.240)	-0.041 (0.127)	-0.472 (0.224)	-0.071 (0.106)	0.616*** (0.157)
Christian	-0.051 (0.038)	0.310 (0.306)	0.510 (0.384)	-0.033 (0.092)	-0.251 (0.426)	-0.004 (0.048)	0.091 (0.129)	0.041 (0.094)	0.191 (0.228)	-0.003 (0.121)	-0.139 (0.213)	-0.075 (0.100)	0.161 (0.149)
Jewish	-0.004 (0.035)	0.067 (0.267)	0.534 (0.336)	0.015 (0.080)	0.393 (0.372)	-0.073 (0.042)	0.116** (0.112)	-0.051 (0.082)	0.388 (0.199)	0.000 (0.105)	-0.139 (0.186)	-0.011 (0.088)	0.095 (0.130)
Obama Supporter	0.025 (0.039)	-0.341 (0.302)	-0.102 (0.380)	0.146 (0.091)	-0.255 (0.421)	-0.004 (0.048)	-0.160 (0.127)	0.178 (0.093)	0.289 (0.225)	-0.021 (0.119)	-0.022 (0.210)	-0.033 (0.099)	-0.051 (0.147)
Trump Supporter	-0.026 (0.041)	-0.141 (0.320)	0.274 (0.402)	-0.085 (0.096)	-0.851 (0.446)	0.054 (0.051)	-0.413** (0.135)	0.191 (0.099)	-0.132 (0.239)	0.189 (0.126)	0.079 (0.223)	-0.068 (0.105)	-0.121 (0.156)
Republican	0.019 (0.038)	0.232 (0.290)	0.108 (0.365)	-0.088 (0.087)	-0.212 (0.404)	-0.007 (0.046)	-0.142 (0.122)	-0.031 (0.090)	-0.284 (0.217)	0.179 (0.114)	0.036 (0.202)	-0.138 (0.095)	-0.073 (0.142)
Constant	4.183 (0.058)	2.382 (0.468)	1.742 (0.589)	0.247 (0.140)	1.911 (0.653)	0.328 (0.074)	1.363 (0.197)	1.245 (0.145)	0.490 (0.350)	0.187 (0.185)	0.809 (0.326)	0.374 (0.154)	0.423 (0.229)

Note: The values in each cell represent the estimated coefficient, with the standard error provided in parentheses below. Bolded coefficients indicate statistical significance at the 90% confidence level. The asterisks (*, **, ***) denote the levels of statistical significance as follows: *: $p < 0.05$ (significant at the 95% confidence level), **: $p < 0.01$ (significant at the 99% confidence level), ***: $p < 0.001$ (significant at the 99.9% confidence level).

Table 13: Likelihood Across Different Game Settings

	Complete Information Game									Incomplete Information Game			
	Competitive Games			Cooperation Games			Mixed-motive Games			Seq. Games	Bayesian Games		Signaling Games
	BL	HS	LS	BL	HP	AP	BL	H-Pun	L-Pun		p=0.5	p=0.9	
GPT-4o	-1.498	-1.363	-1.932	-1.386	-1.386	-1.368	-0.635	-0.646	-1.129	-0.693	-1.386	-1.386	-1.045
GPT-o1	-1.849	-2.003	-1.020	-1.086	-1.304	-1.160	-0.023	-0.023	-0.801	-1.368	-1.384	-1.386	-0.742
GPT-o3-mini	-1.427	-1.066	-1.169	-0.708	-1.253	-0.023	-0.023	-0.023	-1.336	-1.256	-0.708	-1.332	-0.780
Gemma-V2	-2.045	-1.484	-1.818	-1.386	-1.386	-1.386	-1.375	-1.379	-1.326	-1.386	-1.386	-1.386	-0.675
Gemini-2.0	-1.603	-1.663	-1.623	-1.384	-1.370	-0.650	-0.023	-0.023	-0.965	-1.365	-1.366	-1.386	-1.099
Claude-3-Opus	-2.197	-2.197	-2.194	-1.386	-1.177	-0.785	-0.490	-0.023	-0.934	-1.386	-1.386	-1.386	-0.657
Granite-3.1-8B	-2.164	-2.197	-2.135	-1.386	-1.386	-1.372	-1.386	-1.386	-1.386	-1.386	-1.370	-1.386	-2.197
Granite-3.1-MoE	-1.945	-1.897	-1.901	-0.861	-1.280	-0.367	-0.454	-0.478	-0.303	-1.162	-0.752	-1.364	-2.197
LLaMA-3.1-405B	-1.670	-1.819	-2.154	-1.386	-1.386	-1.386	-1.381	-1.373	-1.386	-1.386	-1.386	-1.386	-0.693
LLaMA-3.1-8B	-2.197	-2.197	-2.197	-1.386	-1.386	-1.386	-1.367	-1.382	-1.373	-1.386	-1.386	-1.386	-0.996
InternLM-V2	-2.164	-2.197	-2.197	-1.386	-1.386	-1.368	-1.386	-1.339	-1.386	-1.386	-1.386	-1.386	-0.903
QwQ-32B	-2.162	-1.884	-2.104	-1.194	-1.359	-0.841	-1.273	-0.987	-0.638	-1.281	-1.330	-1.216	-1.080
GLM-4	-2.197	-2.197	-2.197	-1.386	-1.386	-1.386	-1.386	-1.386	-1.386	-1.386	-1.386	-1.386	-0.871
DeepSeek-V2.5	-2.197	-2.197	-2.197	-1.386	-1.386	-1.386	-1.386	-1.386	-1.386	-1.386	-1.386	-1.386	-1.078
DeepSeek-V3	-2.195	-1.683	-2.104	-0.708	-1.253	-0.023	-0.023	-0.023	-1.336	-1.349	-1.086	-1.323	-0.876
DeepSeek-R1	-2.148	-2.111	-2.193	-1.386	-1.386	-1.386	-1.386	-1.385	-1.277	-1.386	-1.386	-1.386	-0.758

Note: Abbreviations: HS = High Stake, LS = Low Stake, BL = Baseline, HP = High Payoff, AP = Asymmetric Payoff, H-Pun = High Punishment, L-Pun = Low Punishment.

C Game Library Design

In this manuscript, we have developed and collected multiple games, the payoff matrices are below. The complete information includes:

Table 14: Competitive Games

(a) Base	(b) High-stake	(c) Low-stake																											
<table border="1"> <tr><td>10, -10</td><td>0, 5</td><td>-5, 8</td></tr> <tr><td>-10, 10</td><td>5, 0</td><td>8, -5</td></tr> <tr><td>0, 0</td><td>5, -5</td><td>-5, 5</td></tr> </table>	10, -10	0, 5	-5, 8	-10, 10	5, 0	8, -5	0, 0	5, -5	-5, 5	<table border="1"> <tr><td>20, -20</td><td>0, 10</td><td>-10, 15</td></tr> <tr><td>-20, 20</td><td>10, 0</td><td>15, -10</td></tr> <tr><td>0, 0</td><td>10, -10</td><td>-10, 10</td></tr> </table>	20, -20	0, 10	-10, 15	-20, 20	10, 0	15, -10	0, 0	10, -10	-10, 10	<table border="1"> <tr><td>3, -3</td><td>0, 1</td><td>-1, 2</td></tr> <tr><td>-3, 3</td><td>1, 0</td><td>2, -1</td></tr> <tr><td>0, 0</td><td>1, -1</td><td>-1, 1</td></tr> </table>	3, -3	0, 1	-1, 2	-3, 3	1, 0	2, -1	0, 0	1, -1	-1, 1
10, -10	0, 5	-5, 8																											
-10, 10	5, 0	8, -5																											
0, 0	5, -5	-5, 5																											
20, -20	0, 10	-10, 15																											
-20, 20	10, 0	15, -10																											
0, 0	10, -10	-10, 10																											
3, -3	0, 1	-1, 2																											
-3, 3	1, 0	2, -1																											
0, 0	1, -1	-1, 1																											

Table 15: Cooperation Games (Stag Hunt)

(a) Base	(b) High-payoff	(c) Asymmetric-payoff												
<table><tr><td>8, 8</td><td>0, 7</td></tr><tr><td>7, 0</td><td>5, 5</td></tr></table>	8, 8	0, 7	7, 0	5, 5	<table><tr><td>20, 20</td><td>0, 7</td></tr><tr><td>7, 0</td><td>5, 5</td></tr></table>	20, 20	0, 7	7, 0	5, 5	<table><tr><td>12, 8</td><td>0, 7</td></tr><tr><td>7, 0</td><td>5, 5</td></tr></table>	12, 8	0, 7	7, 0	5, 5
8, 8	0, 7													
7, 0	5, 5													
20, 20	0, 7													
7, 0	5, 5													
12, 8	0, 7													
7, 0	5, 5													

Table 16: Mixed-Motive Games (Prisoner's Dilemma)

(a) Base	(b) High-punishment	(c) Low-punishment												
<table><tr><td>3, 3</td><td>0, 5</td></tr><tr><td>5, 0</td><td>1, 1</td></tr></table>	3, 3	0, 5	5, 0	1, 1	<table><tr><td>10, 10</td><td>0, 15</td></tr><tr><td>15, 0</td><td>-5, 5</td></tr></table>	10, 10	0, 15	15, 0	-5, 5	<table><tr><td>3, 3</td><td>0, 4</td></tr><tr><td>4, 0</td><td>2, 2</td></tr></table>	3, 3	0, 4	4, 0	2, 2
3, 3	0, 5													
5, 0	1, 1													
10, 10	0, 15													
15, 0	-5, 5													
3, 3	0, 4													
4, 0	2, 2													

1. **Competitive Games:** These games model adversarial interactions where one player's gain is exactly balanced by the other player's loss. The game consists of three variations:
 - Baseline: A standard competitive game where players choose strategies to maximize their own payoff while minimizing their opponent's.
 - High-Stake: A version where payoff differences are significantly larger, increasing the risk and reward of each decision.
 - Low-Stake: A variant where payoff differences are minimized, reducing the overall impact of each decision and testing an agent's ability to navigate lower-risk scenarios.
2. **Cooperation Games (Stag Hunt):** These games examine the trade-off between individual risk and collective reward, highlighting the challenge of trust and cooperation. Three variations are included:
 - Baseline: The classic Stag Hunt game, where mutual cooperation leads to the highest payoff, but unilateral deviation results in significant losses.
 - High-Payoff: A modified version where the rewards for successful cooperation are increased, testing whether agents are more willing to take cooperative risks.
 - Asymmetric-Payoff: A variant where one player receives a higher payoff for cooperation than the other, introducing an imbalance that challenges fairness and trust dynamics.
3. **Mixed-Motive Games (Prisoner's Dilemma):** These games explore the tension between cooperation and self-interest, where defection offers short-term individual benefits but harms collective outcomes. Variations include:
 - Baseline: The standard Prisoner's Dilemma setup, where mutual cooperation yields moderate rewards, but unilateral defection offers a higher individual payoff at the expense of the other player.
 - High-Punishment: A version where the penalty for defection is significantly increased, discouraging selfish behavior.
 - Low-Punishment: A variant where the cost of defection is minimal, encouraging more frequent defection and testing an agent's ability to recognize long-term cooperation benefits.
4. **Sequential Games:** These games test strategic planning and reaction-based decision-making by introducing turn-based interactions.
 - Player 1 makes the first move, setting the stage for Player 2's decision.

Table 17: Sequential Games

0, 5	0, 3	0, 0
5, 2	3, 3	-1, -1
2, 4	4, 3	0, -2

- The full payoff matrix is shown to Player 1, giving them a strategic advantage in planning their move.
- Player 2 must decide based on the observed action of Player 1, testing their ability to infer intent and optimize their response.

In our experiment, only Player 1's responses are collected for strategic reasoning evaluation.

Table 18: Incomplete Information Games Payoff Matrices

(a) Bayesian Coordination Games		(d) Signaling Games																	
(b) Type 1	(c) Type 2	(e) Sender Payoffs	(f) Receiver Payoffs																
<table><tr><td>10, 10</td><td>5, 2</td></tr><tr><td>7, 5</td><td>3, 3</td></tr></table>	10, 10	5, 2	7, 5	3, 3	<table><tr><td>8, 8</td><td>6, 3</td></tr><tr><td>5, 4</td><td>2, 2</td></tr></table>	8, 8	6, 3	5, 4	2, 2	<table><tr><td>5, 5</td><td>2, 1</td></tr><tr><td>3, 2</td><td>1, 0</td></tr></table>	5, 5	2, 1	3, 2	1, 0	<table><tr><td>4, 4</td><td>6, 3</td></tr><tr><td>2, 3</td><td>1, 2</td></tr></table>	4, 4	6, 3	2, 3	1, 2
10, 10	5, 2																		
7, 5	3, 3																		
8, 8	6, 3																		
5, 4	2, 2																		
5, 5	2, 1																		
3, 2	1, 0																		
4, 4	6, 3																		
2, 3	1, 2																		

The incomplete information includes:

1. **Bayesian Coordination Games:** These games test an agent's ability to make optimal decisions under uncertainty. Each round presents two payoff matrices, one of which is randomly selected based on a given probability distribution. The player must make a decision without knowing which matrix is active, relying solely on probabilistic reasoning. The challenge lies in balancing risk and expected utility while considering the likelihood of each scenario. This game type evaluates how well an agent can infer optimal strategies from incomplete information.
2. **Signaling Games:** These games assess asymmetric information processing and strategic signaling. Player 1, the sender, has access to both a real and a fake payoff matrix. Player 2, the receiver, is only given the fake matrix and is aware that it does not reflect the true payoffs. Player 1 can send a signal to influence Player 2's decision, and Player 2 must determine whether to trust or disregard the signal. This game tests an agent's ability to strategically communicate and interpret signals in an environment where deception and trust play key roles.

Table 19: S-W 10

47, 47	51, 44	28, 43
44, 51	11, 11	43, 91
43, 28	91, 43	11, 11

S-W 10 matrix is the matrix we borrow from the human research in behavioral economics in (54; 51).

D Prompt Design

In this section, we present the complete prompts used in our experiments. Three types of experiments were conducted in all cases.

The first type involves the vanilla version of the prompts, with specific variations for the Sequential Game, Bayesian Coordination Game, and Signaling Game, as detailed in Appendix C. The second type includes zero-shot CoT prompts, which add two additional sentences to the base prompts.

For the demographic feature-embedded prompts, a template is followed, where demographic information is incorporated either as part of the system context or as an initial section of the user content, explicitly stating the user’s profile with demographic features.

D.1 Vanilla Experiment Prompt

The prompts used for the baseline evaluation of LLMs are listed in Table 20, 21, 22, and 23. These prompts are repeatedly sent using the LLMs’ API or cloud services’ API.

D.2 Demographic Feature Embedding

The prompts used in the demographic-feature embedded experiment consist of two parts. The main body of the questions remains consistent with those in the vanilla prompt design. However, each prompt is augmented with a demographic component based on the template provided in Table 24. This demographic information is added at the beginning of each prompt to ensure that the LLM retains all features throughout the interaction, preventing memory loss during long-term conversations.

D.3 Chain-of-Thought Embedding

The prompts used in the CoT embedding experiment are straightforward. Following the zero-shot CoT approach, a few additional sentences are appended to the end of each prompt to activate the CoT feature and enhance the reasoning capability of the model. Examples of the added CoT prompts are provided in Table 25.

Table 20: User Prompt for Competitive, Cooperation, and Mixed-motive Game

User Prompt
You are an assistant who helps to make a choice in a game. You should make choices based on the idea of maximizing the number you get from a matrix. Now you are player one/two. You have to pick a row/column number x/y from row 0, 1, and 2. Then there will be a player two/one who will pick column number y/x from column 0, 1, and 2. Your payoff will be the first/second value in location (x, y). Assume the matrix is <A Matrix> Please only give me a result of the row number you picked, do not include any thinking process.

Table 21: User Prompt for Bayesian Coordination Game

User Prompt
You are an assistant who helps to make a choice in a game. You should make choices based on the idea of maximizing the number you get from a matrix. Now you are player one/two. You have to pick a row/column number x/y from row 0 and 1. Then there will be a player two/one who will pick column/row number y/x from column 0 and 1. Your payoff will be the first/second value in location (x, y). With a <p> percent chance, you will be facing Matrix: <A Matrix> . With a <1-p> percent chance, you will be facing Matrix: <B Matrix> . Please only give me a result of the row number you picked, do not include any thinking process.

Table 22: User Prompt for Sequential Game

User Prompt
<p>Now you are player one. You are the first player to pick. You have to pick a row number x from row 0, 1, and 2. Then there will be a player two who will pick column number y from column 0, 1, and 2 based on your decision. Your payoff will be the first value in location (x, y). Assume the matrix is <A Matrix>. Please only give me a result of the row number you picked, do not include any thinking process.</p>

Table 23: User Prompt for Signaling Game

User Prompt	
<p>Player one</p> <p>You should make choices based on the idea of maximizing the number you get from a matrix. Now you are player one. You have to pick a row number x from row 0 and 1. Then there will be a player two who will pick column number y from column 0 and 1. Your payoff will be the first value in location (x, y). The true matrix that determines the payoff is Matrix: <Matrix True>. However, the matrix player two will be seeing is Matrix: <Matrix Fake>. Please only give me a result of the row number you picked, do not include any thinking process.</p>	<p>Player two</p> <p>You should make choices based on the idea of maximizing the number you get from a matrix. Now you are player two. You have to pick a column number y from column 0 and 1. Then there will be a player one who will pick row number x from row 0 and 1. Your payoff will be the second value in location (x, y). The matrix you will be seeing is different from the true matrix, but you have to make your best selection based on your guess and the matrix you see. The matrix is <Matrix Fake>. Please only give me a result of the column number you picked, do not include any thinking process.</p>

Table 24: Demographic Feature Prompt Template

Template
<p>Imagine a ['Age'] year old ['Gender'] with a ['Education'] degree, who is ['Marital Status'] and lives in a ['Location'] area. This individual identifies as ['Sexual Orientation'] and is ['Disability'], of ['Race'] descent, adheres to ['Religion'] beliefs, and supports ['Political Affiliation'] policies. Consider the risk preferences and decision-making processes of a person with these characteristics.</p>

Table 25: Chain-of-Thought version Prompt of Competitive Games

User Prompt
<p>You are an assistant who helps to make a choice in a game. You should make choices based on the idea of maximizing the number you get from a matrix. Now you are player one/two. You have to pick a row/column number x/y from row 0, 1, and 2. Then there will be a player two/one who will pick column number y/x from column 0, 1, and 2. Your payoff will be the first/second value in location (x, y). Assume the matrix is <A Matrix> To decide, analyze the possible outcomes based on maximizing your payoff. Explain your reasoning step by step and then provide only the row/column number you picked as the final answer.</p>

E Complete Reasoning Chains of Claude-3-opus and GPT-4o

Claude-3-opus

Let me understand the situation and analyze it step by step. First, I need to interpret the matrix you've provided. Let me organize it for clarity: Column 0 Column 1 Column 2 Row 0 (10,-10) (0,5) (-5,8) Row 1 (-10,10) (5,0) (8,-5) Row 2 (0,0) (5,-5) (-5,5) From your description, I understand that I am player one and need to pick a row number (x). Then player two will pick a column number (y). My payoff will be the first value in the location (x, y). Since my goal is to maximize my payoff, I need to consider how player two will respond to each of my choices. Let's analyze:

If I choose Row 0:

If player two chooses Column 0, my payoff is 10 If player two chooses Column 1, my payoff is 0 If player two chooses Column 2, my payoff is -5 Since player two wants to minimize my payoff, they would choose Column 2, giving me -5

If I choose Row 1:

If player two chooses Column 0, my payoff is -10 If player two chooses Column 1, my payoff is 5 If player two chooses Column 2, my payoff is 8 Player two would choose Column 0, giving me -10

If I choose Row 2:

If player two chooses Column 0, my payoff is 0 If player two chooses Column 1, my payoff is 5 If player two chooses Column 2, my payoff is -5 Player two would choose Column 2, giving me -5

Looking at the worst outcomes from each of my choices:

Row 0: -5 Row 1: -10 Row 2: -5

I should pick the row that gives me the maximum value among these worst-case scenarios. Both Row 0 and Row 2 give me -5 in the worst case, which is better than Row 1's worst case of -10. Between Row 0 and Row 2, I would choose Row 0 because if player two makes a mistake, I have the potential to get 10 (the highest possible payoff in the entire matrix), whereas with Row 2, the best I could get is 5.

Wait, I made a mistake in my analysis. Let me reconsider this problem using the concept of maximin strategy, which is appropriate for this game. For each row, I need to find the minimum payoff I could receive:

Row 0: $\min(10, 0, -5) = -5$ Row 1: $\min(-10, 5, 8) = -10$ Row 2: $\min(0, 5, -5) = -5$

I should choose the row with the maximum of these minimum values. Both Row 0 and Row 2 have a minimum of -5, which is better than Row 1's minimum of -10. Since Row 0 and Row 2 are tied in terms of their minimum values, I'll look at their potential for higher payoffs. Row 0 has a potential maximum of 10, while Row 2 has a maximum of 5. Therefore, I would choose Row 0.

ORetryClaude does not have the ability to run the code it generates yet. Claude can make mistakes. Please double-check responses.

GPT-4o

We are Player Two (Column chooser) We pick a column y, then Player One picks a row x.

We get the second number from the chosen cell (x, y).

Now we assume Player One picks the row that minimizes our payoff, so again we apply maximin, looking at minimum second values per column:

Column 0: (0,0): -10 (1,0): 10 (2,0): 0 \rightarrow Min = -10

Column 1: (0,1): 5 (1,1): 0 (2,1): -5 \rightarrow Min = -5

Column 2: (0,2): 8 (1,2): -5 (2,2): 5 \rightarrow Min = -5

So: Col 0: -10 Col 1: -5 Col 2: -5 \rightarrow Maximin is -5, achieved by Col 1 or 2

Again, break tie with average: Col 1 values: 5, 0, -5 \rightarrow average = 0 Col 2 values:
8, -5, 5 \rightarrow average = 2.67
So Column 2 is better.

NeurIPS Paper Checklist

1. Claims

Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

Answer: [\[Yes\]](#)

Justification: The abstract and introduction clearly state the claims made.

Guidelines:

- The answer NA means that the abstract and introduction do not include the claims made in the paper.
- The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
- The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
- It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. Limitations

Question: Does the paper discuss the limitations of the work performed by the authors?

Answer: [\[Yes\]](#)

Justification: The paper includes a limitation part to discuss the limitation and future work.

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. Theory assumptions and proofs

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [\[Yes\]](#)

Justification: The paper includes complete proof for the mathematical framework used in the paper.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.

- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. Experimental result reproducibility

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: The information provided in the paper, including prompts, TQRE framework, and the game design, is enough to reproduce the experimental results.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general, releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
 - (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
 - (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
 - (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
 - (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. Open access to data and code

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: Code for experiments is provided in the supplementary material.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so “No” is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.

- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. Experimental setting/details

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [\[Yes\]](#)

Justification: The paper includes all details in the main text.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. Experiment statistical significance

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [\[Yes\]](#)

Justification: The experiment setup in this paper provides enough information to prove the statistical significance of the experiments.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. Experiments compute resources

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [\[No\]](#)

Justification: The experiments are lightweight and not resource-intensive. Reproduction can be achieved on standard hardware without specialized infrastructure.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.

- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. Code of ethics

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics <https://neurips.cc/public/EthicsGuidelines>?

Answer: [Yes]

Justification: Research in the paper conforms with the NeurIPS Code of Ethics.

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. Broader impacts

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: This paper includes a paragraph to discuss the broader impacts.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. Safeguards

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: No data or model in this paper needs to be safeguarded.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.

- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. Licenses for existing assets

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [NA]

Justification: The paper does not use existing assets.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, paperswithcode.com/datasets has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. New assets

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [Yes]

Justification: Code released in this paper is included with the readme file to reproduce or reuse.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. Crowdsourcing and research with human subjects

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: The paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. Institutional review board (IRB) approvals or equivalent for research with human subjects

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: The paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [Yes]

Justification: This paper clearly states that the purpose is to evaluate the reasoning depth of LLMs using the proposed method.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (<https://neurips.cc/Conferences/2025/LLM>) for what should or should not be described.