Revista Latinoamericana de Difusión Científica
Volumen 7 – Número 13
Depósito Legal ZU2019000058 - ISSN 2711-0494

Revista Latinoamericana de Difusión Científica

Volumen 7 - Número 13
Julio – Diciembre 2025
Maracaibo – Venezuela

Game On: A Survey of Strategic Reasoning Benchmarks for Large Language Models

DOI: https://doi.org/10.5281/zenodo.15832279

Daniel José Boza Muñoz*

ABSTRACT

Large Language Models (LLMs) are increasingly deployed in tasks requiring sophisticated strategic reasoning. We systematically surveyed recent benchmarks developed to evaluate LLM-based strategic capabilities across cooperative, adversarial, and policy-oriented contexts. Following a reproducible search strategy, we identified 7 primary benchmarks from a pool of 573 papers. These benchmarks measure diverse dimensions such as multi-step planning, hidden-information inference, cooperative behavior, and deceptive tactics. Their methodologies include multi-agent competitions, board games, negotiation settings, and public goods scenarios, each with distinct metrics—ranging from Elo ratings to Bradley-Terry models—yielding crucial insights into LLM performance. We highlight current limitations, including the lack of standardized metrics and limited real-world applicability, and propose directions for future research, such as human-in-the-loop evaluations and policy-level simulations. Our survey aims to inform researchers and practitioners seeking robust frameworks for assessing LLMs' strategic reasoning proficiencies.

KEYWORDS: Large Language Models, Strategic Reasoning, Benchmarks, Multi-Agent Systems.

*Laboratorio Venezolano de Inteligencia Artificial, Universidad Católica Andrés Bello. ORCID: https://orcid.org/0009-0008-4528-156X. E-mail: larezrafael@gmail.com

Recibido: 23/04/2025 Aceptado: 18/06/2025

Juego digital: Un estudio de los parámetros de razonamiento estratégico para modelos lingüísticos de gran escala

RESUMEN

Los Modelos de Lenguaje de Gran Escala (LLMs) están siendo cada vez más utilizados en tareas que requieren un razonamiento estratégico sofisticado. Realizamos una revisión sistemática de los benchmarks recientes desarrollados para evaluar las capacidades estratégicas de los LLMs en contextos cooperativos, adversariales y orientados a políticas. Siguiendo una estrategia de búsqueda reproducible, identificamos 7 benchmarks principales a partir de un total de 573 artículos. Estos benchmarks miden diversas dimensiones, como la planificación en múltiples pasos, la inferencia de información oculta, el comportamiento cooperativo y las tácticas de engaño. Sus metodologías incluyen competencias multiagente, juegos de tablero, entornos de negociación y escenarios de bienes públicos, cada uno con métricas específicas—desde calificaciones Elo hasta modelos Bradley-Terry—que ofrecen información crucial sobre el desempeño de los LLMs. Destacamos las limitaciones actuales, incluyendo la falta de métricas estandarizadas y la aplicabilidad limitada al mundo real, y proponemos líneas de investigación futura, como evaluaciones con humanos en el bucle y simulaciones a nivel de políticas. Nuestro estudio busca informar a investigadores y profesionales interesados en marcos sólidos para evaluar las competencias de razonamiento estratégico de los LLMs.

KEYWORDS: Modelos lingüísticos de gran escala, Razonamiento estratégico, Parámetros de referencia, Sistemas multiagente.

Introduction

This paper addresses a research gap in the rapidly evolving domain of LLM-centric strategic reasoning by offering a consolidated appraisal of the most pertinent and publicly available benchmarks. While recent years have seen growing interest in leveraging LLMs for tasks that require strategic reasoning, and although numerous surveys cover a range of LLM-related work, there remains a paucity of comprehensive reviews focused specifically on strategic reasoning benchmarks that collate the key studies, assess their methodological rigor, and synthesize their contributions in a manner that is both accessible and replicable.

The increasing adoption of Large Language Models (LLMs) in a wide range of applications—from conversational agents to policy advising—has highlighted the critical

need for systematic evaluations of their strategic reasoning capabilities. Strategic reasoning, broadly defined, involves decision-making processes under conditions of complexity, competition, or cooperation, where agents must anticipate and respond to the actions of other agents or environmental constraints. (Gandhi et al., 2023; Abdelnabi et al., 2024; Duan et al., 2024; Zhang et al., 2024).

Within the context of LLMs, strategic reasoning transcends mere linguistic proficiency and delves into how effectively these models can plan, predict, adapt, and learn from dynamic and often adversarial scenarios.

Building on this understanding, the purpose of the present survey is twofold. First, it seeks to offer a broad overview of existing benchmarks that measure or otherwise characterize the strategic reasoning abilities of LLMs. Second, it aims to serve as a rapid reference point for researchers interested in exploring or expanding upon the current state of the art. To achieve these objectives, the work is structured into six parts: (1) an Introduction to contextualize the problem space and outline the motivation for this survey; (2) a Definition of Strategic Reasoning in the Context of LLMs, which clarifies the core conceptual underpinnings and highlights the nuances of strategic reasoning tasks; (3) a discussion of the Importance of Benchmarks for Evaluating Strategic Reasoning in LLMs, underscoring the role of standardized testbeds for progress in this field; (4) the Parameters of Search and Results of the Systematic Review, providing an in-depth account of our reproducible search strategy and the filtering process; (5) Summarized Analyses of the Identified Benchmarks, offering insights into how each benchmark conceptualizes strategic reasoning, the specific metrics used, and the public repositories available; and (6) Conclusions and Recommendations, where we synthesize key findings, discuss implications, and propose future research directions.

Our methodology hinges on an extensive review of the literature across both arXiv and Google Scholar, capturing published and preprint studies from 2022 to 2024. The search strategy yielded 573 potential papers—225 from arXiv and 348 from Google Scholar. After applying stringent inclusion criteria, 52 relevant articles were selected for closer examination, and 7 studies ultimately met all four of the following criteria:

1. Conversational Games: We prioritized benchmarks that incorporate conversational games. Such environments are highly suitable for evaluating strategic reasoning because they demand a spectrum of capabilities—natural language

understanding, planning under uncertainty, collaborative or adversarial interaction, and real-time adaptability.

2. Open Code Repositories: We required that each benchmark offer a publicly available codebase, ensuring transparency and enabling reproducible studies. The availability of code not only promotes validation of results but also fosters further innovation and critique within the research community.

3. Defined Characteristics of Strategic Reasoning: Each benchmark had to explicitly outline the aspects of strategic reasoning being tested—e.g., negotiation skills, cooperative tactics, adversarial strategies—thereby allowing for targeted and systematic evaluation.

4. Well-Defined Evaluation Metrics: Finally, the benchmark had to provide clearly specified quantitative or qualitative metrics, enabling robust assessment of an LLM's performance vis-à-vis defined strategic reasoning capabilities.

By integrating insights from a carefully curated selection of benchmarks, we endeavor to capture the state of the art and to highlight fertile areas for future research—especially in unexplored contexts such as public policy and multi-agent governance, where the literature remains scant.

1. Definition of Strategic Reasoning in the Context of LLMs

In this survey, we have chosen to adopt the definition of strategic reasoning proposed by Gandhi, Sadigh, and Goodman (2023) because it captures the multifaceted nature of interactions among LLM agents, especially in contexts where objectives may diverge and information is always incomplete. Their conceptualization aligns well with the practical demands of LLMs, which operate not only as language processors but also as decision-makers in dynamic, multi-agent settings.

1.1. Core Aspects of the Adopted Definition

Gandhi et al. (2023) characterize strategic reasoning as the ability of an agent to anticipate and reason about other agents' actions, intentions, and responses within environments where objectives may conflict. Specifically, the authors highlight three interconnected components of strategic reasoning, each of which has direct implications for how LLMs can (and should) be evaluated:

- *Search through the Space of States and Actions:* Agents systematically explore possible actions and resulting states, gauging how these may affect both the environment and other participants. In LLM contexts, this could involve generating various action pathways (e.g., different negotiation offers or dialogue strategies) to determine which approach is most likely to yield desirable outcomes.
- *Assign Values to States and Actions:* Effective strategy hinges on assessing the utility or payoff of different scenarios for oneself and for other agents. By quantifying trade-offs, LLMs can better model the complexities of multi-agent interactions—such as balancing cooperation and competition—while managing uncertainties about others' goals.
- *Form Beliefs about the Partially Observable World:* Agents rarely have full information, necessitating inferences about hidden states or other agents' motivations. LLMs, equipped with sophisticated language understanding, can draw on textual cues to update beliefs as they gather new information, enabling more adaptive strategies over time.

2. Importance of Benchmarks for Evaluating Strategic Reasoning in LLMs

Benchmarks in natural language processing (NLP) and related fields serve as critical tools for measuring progress, ensuring reproducibility, and guiding future research (Zhang et al., 2024; Jiang et al., 2024). Over the past decade, a wide variety of well-established benchmarks have emerged to evaluate language models on diverse tasks such as language understanding, text generation, and question answering. However, recent work underscores that while numerous benchmarks exist to assess an LLM's linguistic capabilities, there is presently no standardized testbed dedicated to evaluating how well these models perform when placed in scenarios that require complex strategic planning and negotiation (Gandhi et al., 2023).

Traditional NLP benchmarks may inadvertently conflate or overlook the nuanced set of cognitive skills inherent in strategic reasoning, including dynamic decision-making, long-horizon planning, belief revision under uncertainty, and adaptability in the face of shifting objectives (Zhang et al., 2024).

Establishing specialized benchmarks for strategic reasoning in LLMs is therefore imperative. Standardized metrics and shared datasets would not only facilitate fair comparisons among different models and approaches but also foster a deeper

understanding of the specific competencies required to excel in strategic environments (Jiang et al., 2024). Moreover, creating a common frame of reference can spur researchers to explore novel modeling techniques that specifically target and refine strategic capabilities in large language models (Zhang et al., 2024).

In addition, the lack of standardized benchmarks for strategic reasoning hinders the formation of consensus regarding best practices (Gandhi et al., 2023). By offering a consolidated overview of the most relevant existing efforts, this survey aims to promote a collective understanding of strategic reasoning tasks and the methodological challenges involved. Through the identification and classification of current benchmarks, coupled with a discussion of their strengths and limitations, this paper seeks to catalyze community-wide dialogue and collaboration. Ultimately, such coordination is expected to encourage the emergence of broadly accepted evaluation protocols and to inspire new benchmarks that can more systematically assess strategic interactions within multi-agent and policy-oriented contexts (Jiang et al., 2024).

3. Search Parameters and Systematic Review Outcomes

This section outlines the systematic search parameters employed across two primary sources—arXiv and Google Scholar—and details the resulting pool of articles. The objective was to identify research studies addressing strategic reasoning in the context of Large Language Models (LLMs). All searches were designed to be reproducible, with clear inclusion and exclusion criteria. Below, we provide a detailed account of each search, followed by the total number of documents retrieved, screened, and ultimately included in the final analysis.

3.1. Searches in arXiv

Four separate searches were conducted (after trial and error) within the arXiv repository, each focusing on different keyword combinations and date ranges. The overall aim was to capture articles related to strategic reasoning in LLMs, while excluding documents that centered on code generation or otherwise fell outside the scope of the survey.

- *First Search*
    - Title: “reasoning” AND NOT “code”
    - Abstract: “large language model” AND “strategy”
    - Date Range: January 1, 2022 – October 12, 2024
    - Outcome: 164 articles retrieved
    - Rationale: This search targeted works on strategic reasoning in large language models, explicitly excluding those focused on code generation. It yielded 164 articles, proving to be the most effective search in terms of identifying a substantial number of relevant documents (ideal range: 100–500).
- *Second Search*
    - Title: “strategic” AND “reasoning”
    - Abstract: “agent” AND “llm” AND “evaluating”
    - Date Range: 2022–present
    - Outcome: 2 articles retrieved
    - Rationale: This was a narrower search, concentrating on evaluating strategic agents in the context of LLMs. The highly specific criteria led to a limited set of only two articles.
- *Third Search*
    - Title: “strategic” AND “reasoning”
    - Abstract: “public policy”
    - Filter: Computer Science
    - Date Range: 2022–present
    - Outcome: 0 articles retrieved
    - Rationale: No articles were found that directly connected strategic reasoning, public policy, and LLMs within the Computer Science category. This result underscores a potential gap in the literature at the intersection of strategic reasoning in LLMs and public policy.
- *Fourth Search*
    - Abstract: abstract=“Large Language Models” AND abstract=“strategic reasoning” AND abstract=“government”
    - Outcome: 3 articles retrieved

- Rationale: This search targeted studies relating large language models, strategic reasoning, and governmental contexts. Although it yielded a small number of results, it highlights an emerging or relatively unexplored research area.

- *Fifth Search*

- Abstract: abstract="Large Language Models" AND abstract="public policy"

- Outcome: 56 articles retrieved

- Rationale: This final arXiv search sought works integrating large language models with public policy. The moderate number of articles (56) suggests a developing domain of study connecting LLMs to policy-related inquiries.

3.2. Searches in Google Scholar

Three principal searches were conducted using Google Scholar. These aimed to complement the arXiv results by capturing a broader cross-section of peer-reviewed work, conference papers, and other scholarly outputs. Each search employed Boolean operators and keywords to narrow the focus to strategic reasoning tasks in LLMs, particularly those involving multi-agent scenarios, benchmarking, and frameworks for evaluation.

- *First Search*

- Expression: (llm AND reasoning AND planning AND strategic AND agent AND benchmark AND framework) AND NOT code

- Outcome: 7,840 articles retrieved

- Rationale: Although the search expression was designed to pinpoint articles addressing strategic reasoning, planning, benchmarking, and frameworks in LLMs, it returned an exceedingly broad set of results. Consequently, the majority were screened out during the eligibility assessment.

- *Second Search*

- Expression: ("Large Language Models" OR LLM) AND "strategic reasoning" AND (agent OR agents) AND (benchmark OR evaluating) AND (framework OR model) AND NOT code

- Outcome: 263 articles retrieved

- Rationale: This refined query significantly reduced the number of results from 7,840 to 263 by adding parentheses, explicit logical operators, and additional keywords such as “benchmark,” “evaluating,” and “framework.”

- *Third Search*

- Expression: (“Large Language Models” OR LLM) AND “strategic reasoning” AND (agent OR agents) AND (government OR “public policy”) AND NOT code
- Outcome: 85 articles retrieved
- Rationale: The final Google Scholar search specifically looked for articles bridging large language models, strategic reasoning, and agents within governmental or public policy contexts, excluding code-related works. The resulting 85 articles formed a manageable subset for more detailed review.

3.3. Final Selection

After completing all searches across both arXiv and Google Scholar, a total of 573 articles were initially identified—225 from arXiv and 348 from Google Scholar. Each paper underwent a multi-stage screening process based on predetermined inclusion and exclusion criteria, resulting in 52 articles deemed relevant. Of these, 7 articles fulfilled all four inclusion criteria outlined in the Introduction. These 7 benchmarks form the basis of the in-depth analysis presented in subsequent sections of this survey.

4. Summarized Analyses of the Identified Benchmarks

4.1. GameEval

Methodology

Qiao, D., Wu, C., Liang, Y., Li, J., & Duan, N. (2023) design three distinct games—Ask-Guess, SpyFall, and TofuKingdom—to evaluate a range of strategic reasoning capabilities in LLMs. Implementation involves role-based multi-turn conversations, private history maintenance, and chain-of-thought prompting to simulate realistic interactions.

Strategic Reasoning Capabilities Evaluated

The paper evaluates cooperative and adversarial strategies, specific knowledge, multi-hop reasoning, deceptive strategies, long-term planning, and instruction-following.

1. Cooperative and adversarial strategies reflect real-world social dynamics.
2. Specific knowledge: models ability to apply relevant information in context.
3. Multi-hop reasoning: involves integrating information over multiple steps, which is crucial for handling complex tasks.
4. Deceptive strategies: test model's ability to simulate human-like behaviors in adversarial settings
5. Long-term planning: emphasizes foresight and anticipation.
6. Instruction-following: ensures adherence to defined frameworks and constraints.

Evaluation Metrics

For each game, specific evaluation metrics were established to quantify the performance of the LLMs.

- Ask-Guess

- Successful Trial (ST): The model correctly guesses the word within the limited number of rounds without violating game rules.
- Ending Error (EE): The answerer ends the game prematurely.
- Round Limit Error (RLE): The model fails to guess the word within the maximum allowed rounds (set to 30), indicating inefficiency in reasoning or questioning strategies.
- Answer Mentioned Error (AME): The answerer mentions the secret word directly, violating the game rules.
- Chat Error (CE): Errors due to API request failures or generation issues.
- Average Number of Rounds: For successful trials, the average number of Q&A rounds taken to guess the word, reflects the model's efficiency in narrowing down possibilities.

- SpyFall

- Spy Winning Rate: The proportion of games where the spy avoids detection until the end, indicating the model's effectiveness in deception and maintaining cover under scrutiny.
- Spy Living Rounds: The average number of rounds the spy survives before being identified and eliminated, reflecting the model's ability to sustain deceptive strategies over time.

- TofuKingdom

- Points Earned: The cumulative points across multiple game iterations indicate the model's overall performance in adopting roles, strategizing, and achieving objectives under complex game dynamics.

4.2. GTBench

Methodology

Duan et al. (2024) centers on engaging LLMs in a series of ten well-recognized games that span a wide range of game-theoretic taxonomies—encompassing complete information (Tic-Tac-Toe, Connect-4, Breakthrough), incomplete information (Kuhn Poker, Liar's Dice), dynamic and static interactions (Iterated Prisoner's Dilemma, Blind Auction), and probabilistic (Nim) and deterministic scenarios (Negotiation, Pig). The authors structured their methodology around two primary investigative approaches:

1. Characterizing Strategic Reasoning of LLMs: Comparing their performance to conventional game-solving algorithms and examining the impact of factors such as pre-training, model size, and reasoning methods.

2. LLM-vs.-LLM Competitions as Reasoning Evaluation: the authors implemented LLM-versus-LLM competitions. They employed a modular prompting strategy that consisted of a system prompt, a head prompt, an observation prompt, and a reasoning prompt.

To analyze the reasoning methods employed by the LLMs, the authors tested several prompting paradigms:

1. Direct Prompting: Where the LLM generates responses without additional reasoning steps.

2. Chain-of-Thought (CoT): Encouraging the LLM to think step-by-step before generating an action.

3. Self-Consistent CoT (SC-CoT): Generating multiple reasoning trajectories and using majority voting to decide on the final action.

4. Tree-of-Thought (ToT): Incorporating exploration and self-evaluation in the reasoning process.

Strategic Reasoning Capabilities Evaluated

The paper evaluates the following strategic reasoning traits in Large Language Models (LLMs):

1. Pure Logical and Strategic Reasoning: The ability to engage in reasoning that relies solely on logic and strategy without the influence of complex narratives or character roles.

2. Handling Complete vs. Incomplete Information: The capacity to perform strategic reasoning in games with full visibility of the game state (complete information) and in games where some information is hidden (incomplete information).

3. Dynamic vs. Static Game Reasoning: The ability to make strategic decisions in multi-turn games with evolving states (dynamic) versus single-turn games with fixed states (static).

4. Probabilistic vs. Deterministic Reasoning: The skill to reason and make decisions in environments where outcomes are probabilistic (involving chance) versus deterministic (predictable outcomes based solely on players' actions).

5. Board Strategy Skills: Proficiency in planning and executing strategies in board games that require positional play and foresight.

6. Collaboration and Negotiation Abilities: The capacity to work cooperatively with other agents towards a common goal or to negotiate mutually beneficial outcomes.

7. Auction and Bidding Skills: The ability to engage in competitive scenarios involving bidding strategies and valuation assessments.

8. Bluffing and Deception: Using misinformation or concealment to gain a strategic advantage over opponents.

9. Mathematical Reasoning and Calculations: The ability to perform precise mathematical computations essential for strategic decision-making.

10. Error Detection and Correction: The capability to recognize and rectify mistakes in reasoning or calculations during gameplay.

Evaluation Metrics

The study employed two primary evaluation metrics:

1. Normalized Relative Advantage (NRA): This metric measures the relative advantage of one participant (e.g., an LLM) over another (e.g., a conventional solver or another LLM) in a series of games. See Duan et al. (2024) for details on the formula.

2. Elo Rating System: For zero-sum games, the authors utilized the Elo rating system to calculate the relative skill levels of the LLMs. This system is widely used in competitive games like chess and provides a dynamic rating that reflects a player's performance relative to their opponents.

4.3. GameBench

Methodology

Costarelli et al. (2024) focus on assessing LLMs across nine diverse game environments, it involves selecting “games that are obscure and unlikely to have been significantly represented in the LLMs' pretraining data”. The games chosen include “Air, Land, and Sea” (ALS), “Arctic Scavengers” (ARC), “Are You the Traitor?” (AYT), “Codenames” (CN), “Hive” (HV), “Pit” (PT), “Santorini” (SN), “Two Rooms and a Boom” (TRB), and “Sea Battle” (SB). They assess the models in their base forms and augmented with two scaffolding techniques designed to enhance strategic reasoning:

1. Chain-of-Thought (CoT) Prompting (mentioned before).

2. Reasoning via Planning (RAP): A scaffolding approach where the model engages in planning-based reasoning, predicting potential future states, and making decisions based on these projections.

Agents play matches against each other across the selected games, facilitating a comprehensive assessment of their strategic reasoning abilities in various contexts. The authors also include a random-action baseline and a human baseline.

Strategic Reasoning Capabilities Evaluated

The authors evaluate the following strategic reasoning traits:

1. Abstract Strategy Reasoning: involves the ability to engage in high-level planning and make decisions based on logical deduction and foresight

2. Reasoning under Non-Deterministic Outcomes: This characteristic assesses an agent's ability to make optimal decisions in environments where outcomes are uncertain.

3. Reasoning with Hidden Information: evaluates how agents perform when all variables are not known, requiring inference and hypothesis generation.

4. Language Communication: Effective strategic reasoning often requires communication to coordinate actions and share information.

5. Social Deduction and Bluffing: involves understanding and predicting other agents' intentions and possibly deceiving them for strategic advantage.

6. Cooperation between Players: this trait examines how well they can work with others to achieve common goals, reflecting on their adaptability and social intelligence within strategic contexts.

Evaluation Metrics

The paper employs the exponential Bradley-Terry model (Bradley & Terry, 1952). This probabilistic model estimates the likelihood that one agent will outperform another based on their assigned ratings, which reflect their latent abilities. The authors state that the Bradley-Terry model has advantages over alternatives like the Elo rating system, for example, due to its assumption that each agent's ability is fixed and does not change over time.

4.4. MAgIC

Methodology

Xu et al. (2024) proposes a competition-based benchmark with five scenarios: two social deduction games—Chameleon and Undercover—and three game-theory scenarios—Cost Sharing, Multi-turn Prisoner's Dilemma, and Public Goo. These scenarios are chosen because they encapsulate key characteristics explained below. Furthermore, the authors introduce an enhancement for LLM agents by integrating Probabilistic Graphical Models (PGMs), creating a “PGM-aware agent” This integration aims to augment the LLM’s capacity to comprehend intricate scenarios by incorporating Bayesian statistical foundations.

Strategic Reasoning Capabilities Evaluated

The paper evaluates the following strategic reasoning traits:

1. Judgment: crucial for assessing incomplete or partial information to make accurate decisions under uncertainty.

2. Reasoning: involves logical analysis and multi-hop thinking to understand complex scenarios and predict outcomes based on others' potential actions.

3. Deception: pertinent in competitive settings where misleading others can provide a strategic advantage.

4. Self-awareness: enables agents to understand their own roles and capabilities, ensuring consistent and appropriate behavior within the system.

5. Cooperation: essential for working effectively with others towards shared objectives, highlighting the social aspect of strategic reasoning.

6. Coordination: involves aligning actions and facilitating agreements among multiple parties, which is vital for successful collaboration.

7. Rationality: pertains to making optimal decisions that maximize benefits by logically considering the potential actions of others rather than acting impulsively.

Evaluation Metrics

Each feature is measured or observed through specific scenarios and quantitative metrics:

1. Judgment: It is calculated as the ratio of correct votes or decisions made by the agent based on partial information in games like Chameleon and Undercover..

2. Reasoning: It is calculated by comparing the agent's deductions with the ground truth and the actual subjective deductions of other agents.

3. Deception: Calculated as the ratio of successful deceptions, such as blending in without being detected or causing incorrect secret word guesses.

4. Self-awareness: Measured by the accuracy of the agent's identification of its own role in games with undisclosed roles.

5. Cooperation: In the Cost Sharing game, it is measured by the number of successful collaborations that result in unanimous agreements.

6. Coordination: Measured by the number of successful collaborations proposed by the agent in the Cost Sharing game.

7. Rationality: In Prisoner's Dilemma and Public Good, rationality is measured by the proportion of decisions that optimize the agent's outcomes according to game rules.

4.5. LLM Deliberation Benchmark

Methodology

Abdelnabi et al. (2024) introduce an evaluation framework LLMs in complex negotiation tasks within multi-agent systems. Negotiations encompass five issues with several sub-options, resulting in 720 possible deal combinations. Agents engage in multi-turn negotiations, aiming to maximize their utility while considering others' preferences, necessitating arithmetic calculations, inference, exploration, planning, and theory-of-mind reasoning. Various game variants are introduced—including compromising, greedy, and adversarial games—to evaluate critical safety aspects and the impact of different agent behaviors on negotiation outcomes.

Strategic Reasoning Capabilities Evaluated

The paper evaluates the following strategic reasoning traits:

1. Strategic Planning: The capacity of agents to formulate and adjust strategies over time to achieve negotiation goals.

2. Cooperation: The ability to work collaboratively with other agents to reach mutually beneficial agreements.

3. Competition: Navigating interactions where agents have conflicting interests, aiming to maximize individual gains.

4. Balancing Multiple Objectives: Managing and reconciling multiple, potentially conflicting, goals within the negotiation.

5. Manipulation and Deception Awareness: Recognizing and appropriately responding to manipulation or deception by other agents.

6. Theory-of-Mind (ToM) Capabilities: Understanding and reasoning about the beliefs, desires, and intentions of other agents.

7. Commonsense Reasoning: Applying general world knowledge and common sense to interpret negotiation contexts.

8. Arithmetic Reasoning: Accurately performing calculations necessary for evaluating proposals and assessing their value.

9. Inference: Drawing logical conclusions from partial observations and interaction history.

10. Exploration: Generating and considering alternative strategies and proposals.

11. Multi-turn Reasoning: Sustaining coherent reasoning processes over multiple interaction rounds.

12. Adversarial Thinking: Anticipating and mitigating the impact of adversarial agents within the negotiation.

13. Safety Considerations: Ensuring robustness and alignment in the presence of manipulation and exploitation attempts.

Evaluation Metrics

The study employs several evaluation metrics to quantify LLM agents' performance:

1. Final Success Rate: measures the proportion of games where the final deal satisfies acceptance thresholds of all relevant parties, indicating agents' ability to reach successful agreements by the negotiation's end.

2. Any Success Rate: assesses the agents' capacity to generate acceptable proposals at any point during negotiations, reflecting flexibility.

3. Own Score: evaluates whether agents effectively maximize their utility based on their secret preferences, indicating self-interested strategic behavior.

4. Collective Score: measures how well an agent's proposals accommodate others' preferences, reflecting cooperative efforts.

5. Wrong Deals Rate: calculates the frequency of agents proposing deals that do not meet their own acceptance thresholds, indicating errors in reasoning or calculation.

6. Score Leakage Ratio: assesses agents' ability to maintain confidentiality by measuring the proportion of communications revealing secret information, crucial for realistic negotiations.

4.6. GamaBench

Methodology

Huang et al. (2024) introduce GAMA (γ)-Bench, a framework designed to evaluate the decision-making abilities of Large Language Models (LLMs), selecting eight classical games categorized into Cooperative Games ("Guess 2/3 of the Average," "El Farol Bar," "Divide the Dollar"), Betraying Games ("Public Goods Game," "Diner's Dilemma," "Sealed-Bid Auction"), and Sequential Games ("Battle Royale," "Pirate Game"). The authors explore various conditions, including different temperature settings, prompt templates, and reasoning strategies like CoT.

Strategic Reasoning Capabilities Evaluated

The study assesses several strategic reasoning traits in LLMs:

1. Perception and Understanding of Game Rules: Evaluated by the models' ability to make valid moves within the game's constraints.

2. Theory of Mind Reasoning: Assessed through games like "Guess 2/3 of the Average," where anticipating other agents' choices is crucial.

3. Strategic Planning and Decision-Making: Observed in multi-round games requiring long-term payoff optimization, such as the "Public Goods Game."

4. Cooperation versus Self-Interest: Examined by analyzing choices between collective welfare and individual benefit.

5. Adaptiveness and Learning from History: Measured by the models' ability to adjust strategies based on previous outcomes in games like "El Farol Bar."

6. Critical Thinking and Integration of Information: Required for synthesizing information to make optimal decisions in complex scenarios.

7. Arithmetic and Quantitative Reasoning: Tested in games necessitating calculations, like the "Sealed-Bid Auction."

8. Dealing with Incomplete or Imperfect Information: Evaluated in scenarios without full knowledge of other agents' actions.

9. Sequential Decision-Making: Assessed in games with sequential moves, such as the "Pirate Game."

10. Robustness and Generalizability: Determined by the models' consistent performance across varying game settings.

Evaluation Metrics

1. Score in “Guess 2/3 of the Average”: Evaluates iterative reasoning and adaptation to collective behavior by examining how closely agents converge toward the minimum allowable number.

2. Score in “El Farol Bar” Game: Assesses the model’s capacity to optimize attendance in a congestion scenario by measuring how closely the number of attendees aligns with the bar’s capacity threshold.

3. Score in “Divide the Dollar” Game: Examines fairness and strategic resource allocation by comparing proposed divisions of a fixed resource with the total amount available.

4. Score in “Public Goods Game”: Gauges tendencies toward cooperation or free-riding by tracking average contributions against the equilibrium strategy of contributing nothing.

5. Score in “Diner’s Dilemma” Game: Analyzes how often agents choose the cheaper dish, reflecting whether they adhere to the dominant (but collectively suboptimal) strategy of selecting the expensive option.

6. Score in “Sealed-Bid Auction” Game: Captures strategic bidding behavior by comparing how agents balance winning against the risk of overpaying, indicated by the difference between bids and valuations.

7. Score in “Battle Royale” Game: Evaluates targeting decisions in a competitive setting by determining whether agents select the most threatening opponent to maximize survival probability.

8. Scores in “Pirate Game” (Proposer Score and Voter Score): Measures the optimality of resource-allocation proposals and the accuracy of voting decisions, indicating alignment with strategic best interests in sequential bargaining.

4.7. GLEE

Methodology

Shapira et al. (2024) introduce GLEE (Games in Language-based Economic Environments), a comprehensive framework and benchmark designed to standardize research on two-player, sequential, language-based games involving Large Language Models (LLMs). The framework encompasses three fundamental families of games—bargaining, negotiation, and persuasion—each grounded in classical economic models and consistently parameterized for controlled experimentation across diverse economic contexts. The methodology emphasizes three critical degrees of freedom: game horizon (the number of time periods and whether the length is known), information structure (agents' awareness of each other's preferences), and communication form (natural language or structured messages).

Strategic Reasoning Capabilities Evaluated

The study assesses ten key strategic reasoning traits of LLMs:

1. Rational Decision-Making: is foundational to any strategic agent, reflecting the ability to make choices that maximize expected utility based on available information.

2. Long-Term Planning and Anticipation: involve forecasting future states and actions, crucial for strategies that unfold over multiple interactions or time periods.

3. Dealing with Information Asymmetry: tests an agent's ability to operate under uncertainty and leverage private information.

4. Strategic Communication and Persuasion: assess how agents use language to influence others' beliefs and actions.

5. Fairness Considerations: relate to the agent's propensity to consider equitable outcomes, which can impact long-term cooperation and reputation.

6. Efficiency Optimization: focuses on achieving outcomes where resources are allocated in a manner that maximizes total benefit.

7. Adaptation to Economic Environment Parameters: evaluates the flexibility of agents to adjust strategies based on variables like game horizon, discount factors, and communication forms.

8. Cooperative and Competitive Behavior: examines the balance between self-interest and mutual benefit, essential for navigating interactions that can be zero-sum or positive-sum.

9. Equilibrium Convergence looks at whether agents' interactions stabilize over time into predictable patterns, reflecting strategic equilibrium concepts like Nash Equilibrium.

10. Language-Based Negotiation Tactics: explore how effectively agents use natural language in negotiation contexts to achieve desired outcomes.

Evaluation Metrics

The study employs three primary evaluation metrics: self-gain, efficiency, and fairness. Self-gain measures the individual utility or payoff an agent achieves, reflecting its effectiveness in maximizing personal benefit within the game. Efficiency evaluates the optimality of outcomes for all parties by normalizing the sum of agents' utilities relative to the maximum possible utility, thus indicating how well agents coordinate to achieve mutually beneficial results. Fairness assesses the equity of outcomes by quantifying deviations from an equitable benchmark, such as an equal division in bargaining games. These metrics are consistently applied across the different game families, facilitating comparative analysis of LLM agents' strategic behaviors. The metrics contribute to the overall findings by highlighting how various strategic reasoning capabilities influence performance. For instance, high self-gain may indicate strong competitive strategies, while high efficiency and fairness scores suggest effective cooperation and equitable decision-making.

Conclusions and Recommendations

The benchmarks surveyed in this paper—GameEval, GTBench, GameBench, MAgIC, LLM Deliberation, GAMA-Bench, and GLEE—collectively demonstrate the rapidly expanding efforts to gauge strategic reasoning in Large Language Models (LLMs). Each benchmark highlights a particular slice of the strategic reasoning spectrum, covering capabilities such as long-horizon planning, cooperative and adversarial interactions, dynamic adaptation to incomplete information, and the balancing of multiple objectives. Taken as a whole, these benchmarks underscore three key observations:

1. **Multi-Dimensional Complexity.** Strategic reasoning in LLMs is a multifaceted construct that cannot be fully captured by a single task or metric. Tasks that mix cooperation, competition, partial observability, and open-ended language interaction reveal strengths and gaps in an LLM's reasoning stack.

2. **Growing Emphasis on Naturalistic Interaction.** Several of the identified benchmarks incorporate natural language dialogues and real-time decision-making, pushing beyond static puzzle-solving or purely logical tasks. This trend reflects the field's turn toward more ecologically valid evaluations, mirroring the real-world scenarios where LLMs must collaborate, negotiate, or compete with humans or other artificial agents.

3. **Interplay Between Reasoning Techniques and Performance.** Across all benchmarks, performance often hinges on how prompting paradigms—such as Chain-of-Thought (CoT), Self-Consistency CoT, Tree-of-Thought (ToT), or planning-based scaffolds—shape the LLM's ability to engage in multi-step reasoning. These techniques can significantly affect outcomes by influencing whether models successfully maintain role consistency, handle hidden information, or adapt to adversarial settings.

Despite these advances, clear gaps remain. Most benchmarks focus on small-scale or stylized games, leaving more complex real-world strategic contexts—especially in domains like policy-making, multi-party negotiations, or large-scale societal coordination—relatively unexplored. Moreover, the field has yet to converge on shared standards for evaluation metrics, data collection protocols, and reproducible experimental frameworks.

Limitations of Existing Benchmarks

A number of key limitations emerged from our review:

1. **Limited Real-World Complexity.** While games like SpyFall and negotiation tasks offer partial analogies to real-world strategic interactions, they often simplify the socio-political, regulatory, and ethical constraints that characterize human decision-making.

2. **Constrained Scenario Diversity.** Benchmarks such as GameEval, GTBench, and GAMA-Bench feature diverse games, but many remain within a narrow scope of puzzle- or table-top-like tasks. This restricts the range of skills tested, such as moral decision-making, large-scale resource allocation, or dynamic coalition formation.

3. **Evaluation Metric Inconsistency.** A variety of rating or scoring systems (e.g., Elo, Bradley-Terry, Normalized Relative Advantage) are employed, hindering direct comparisons across studies. Additionally, metrics often lack consensus definitions for key concepts like “success,” “cooperation,” or “efficiency,” making cross-benchmark synthesis challenging.

4. **Limited Transparency in Prompting Strategies.** While authors frequently report prompting schemes (e.g., CoT, self-consistency), the exact architecture of prompts, hyperparameters, and interactive dynamics are sometimes insufficiently documented or standardized. This reduces the reproducibility and interpretability of results.

5. **Underexplored Policy and Governance Contexts.** Despite some inclusion criteria targeting “public policy” and “government,” most filtered papers did not yield robust benchmarks that capture the complexity of strategic reasoning in large-scale societal systems. The notable absence of real-world policy tasks suggests an important avenue for future work.

Recommendations for Future Research

Future research in this domain can benefit from moving beyond stylized or board-game-style settings and incorporating tasks with greater environmental complexity, such as simulations that involve large-scale resource allocation, dynamic coalitions, and evolving socio-political constraints. Open-world or real-time strategy games could offer deeper insights into how LLMs handle unforeseen contingencies and long-term consequences.

Equally important is the standardization of evaluation protocols, which involves developing shared reporting standards and consistent interfaces for gameplay and data collection. Universal metrics—such as cooperation ratios, negotiation efficiency, or social welfare—and transparent documentation of prompt templates, hyperparameter settings, and scenario rationales would enable replicability and fair model comparisons.

Additionally, while AI-vs.-AI benchmarks are informative, they often miss the nuances of human communication styles, emotional cues, and irrational behaviors. Introducing human participants or high-fidelity human simulations could more accurately evaluate an LLM’s capacity for empathetic negotiation, persuasion under uncertainty, and real-world decision-making.

There is also a pressing need to fill the gap in benchmarks focusing on public policy, governance, and large-scale organizational contexts. Examples include legislative simulations (where models craft bills, debate trade-offs, and form coalitions), budget allocation scenarios (reflecting real-world constraints like earmarked funds and political alliances), and tasks requiring regulatory compliance or ethical decision-making.

As strategic interactions grow increasingly complex, transparency and interpretability become paramount. Future work should investigate explainability frameworks—such as causal modeling or Bayesian updates—and perform robustness checks by testing LLM agents against adversarial strategies, biased data, or atypical prompts.

Finally, combining the strengths of prompting-based language approaches with Multi-Agent Reinforcement Learning (MARL) or structured planning modules could yield hybrid models that excel in both flexible communication and sequential decision-making, ultimately fostering more sophisticated strategic capabilities.

References

Abdelnabi, S., Gomaa, A., Sivaprasad, S., Schönherr, L., & Fritz, M. (2024). Cooperation, competition, and maliciousness: LLM-stakeholders interactive negotiation [Preprint]. arXiv. https://arxiv.org/abs/2309.17234

Duan, J., Zhang, R., Diffenderfer, J., Kailkhura, B., Sun, L., Stengel-Eskin, E., Bansal, M., Chen, T., & Xu, K. (2024). GTBench: Uncovering the Strategic Reasoning Limitations of LLMs via Game-Theoretic Evaluations. arXiv preprint arXiv:2402.12348.

Gandhi, K., Sadigh, D., & Goodman, N. D. (2023). Strategic reasoning with language models. arXiv. https://arxiv.org/abs/2305.19165

Huang, J.-t., Li, E. J., Lam, M. H., Liang, T., Wang, W., Yuan, Y., Jiao, W., Wang, X., Tu, Z., & Lyu, M. R. (2024). How far are we on the decision-making of LLMs? Evaluating LLMs' gaming ability in multi-agent environments. *Artificial Intelligence*. arXiv. https://arxiv.org/abs/2403.11807

Jiang, B., Xie, Y., Wang, X., Su, W. J., Taylor, C. J., & Mallick, T. (2024). Multi-modal and multi-agent systems meet rationality: A survey. https://openreview.net/forum?id=9Rtm2gAVjo

Qiao, D., Wu, C., Liang, Y., Li, J., & Duan, N. (2023). GameEval: Evaluating LLMs on Conversational Games. arXiv preprint arXiv:2308.10032.

Shapira, E., Madmon, O., Reichart, R., & Tennenholtz, M. (2024). Can LLMs replace economic choice prediction labs? The case of language-based persuasion games (v4). arXiv:2401.17435. https://arxiv.org/abs/2401.17435

Xu, L., Hu, Z., Zhou, D., Ren, H., Dong, Z., Keutzer, K., Ng, S. K., & Feng, J. (2024). MAgIC: Investigation of large language model powered multi-agent in cognition, adaptability, rationality and collaboration. *arXiv*. https://arxiv.org/abs/2311.08562

Zhang, Y., Mao, S., Ge, T., Wang, X., de Wynter, A., Xia, Y., Wu, W., Song, T., Lan, M., & Wei, F. (2024). LLM as a mastermind: A survey of strategic reasoning with large language models. arXiv:2404.01230. https://arxiv.org/abs/2404.01230

Conflicto de interés

Los autores de este manuscrito declaran no tener ningún conflicto de interés.

Copyright

Licencia CreativeCommons