Strategic Insights: Evaluating Large Language Models' Decision-Making in Multi-Player Game-Theoretic Environments

Anonymous Author(s)
Affiliation
Address
`email`

Abstract

Large Language Models (LLMs) excel in language tasks but their strategic decision-making in interactive, multi-agent scenarios—critical for applications like negotiation systems or social simulations—remains understudied. This paper examines twelve anonymized LLMs in six multi-player game theory scenarios, encompassing cooperative, betraying, and sequential categories, with ten agents per instance across repeated rounds and multiple runs. We propose the Strategic Rationality Score (SRS), a novel composite metric normalizing deviations from Nash equilibria across games, enabling quantitative benchmarking of LLM rationality. Our findings reveal inconsistent equilibrium-seeking behavior, weak correlations with architectural features like parameter size, and minimal adaptation over interactions, suggesting inherent limitations in opponent modeling and long-term reasoning. These results contrast with expectations from scaling laws and highlight biases toward short-term gains. Contributions include SRS for cross-game evaluation, large-scale multi-player simulations (360 instances), and linkages to LLM traits, advancing AI behavioral analysis for safer multi-agent deployments. Data and code are available as *Supplementary Material* (attachment) to this submission, as well as at: `https://anonymous.4open.science/r/Agents4Science_2025_LLM_Game_Theory-PPPP`.

1 Introduction

The evolution of Large Language Models (LLMs) has revolutionized artificial intelligence, enabling unprecedented proficiency in tasks ranging from natural language understanding to creative generation [23]. As these models integrate into dynamic, interactive systems—such as autonomous agents in virtual economies, collaborative robotics, or policy simulations—their ability to make strategic decisions under uncertainty and interdependence becomes paramount [11]. Game theory, with its formal models of rational choice in conflicting or cooperative settings [38], offers a powerful lens to probe LLM behavior beyond static benchmarks [22].

Existing evaluations often limit to dyadic games, like the Prisoner's Dilemma, where LLMs show cooperative tendencies but susceptibility to framing effects and inconsistent rationality [2, 26]. However, real-world applications involve multi-player dynamics ($N > 2$), introducing complexities like coalition formation, free-riding, and sequential planning, which amplify strategic depth and reveal potential biases [35]. For instance, in resource-sharing simulations or bargaining protocols, irrational LLM decisions could propagate inefficiencies or ethical misalignments [27]. This gap motivates our study: a comprehensive analysis of LLM strategic behavior in scaled multi-player games, linking performance to model architectures and inferred cognitive traits.

Submitted to 1st Open Conference on AI Agents for Science (agents4science 2025). Do not distribute.

The significance of this work lies in its implications for AI alignment and societal impact. Understanding LLM deviations from equilibria can inform safer designs, mitigating risks in high-stakes interactions [16]. Moreover, by simulating human-like agents, LLMs could accelerate behavioral economics research, but only if their limitations are characterized [15]. Our contributions enhance this domain through:

- **Large-Scale Multi-Player Evaluation:** Simulating 360 instances with N=10 agents across diverse game categories, extending beyond prior two-player foci [2].
- **Novel Strategic Rationality Score (SRS):** A weighted, normalized metric for aggregating equilibrium deviations, facilitating comparable rationality assessments and predictive modeling.
- **Trait-Linked Insights:** Correlating performance with LLM features (e.g., parameter size, Theory of Mind inferences), revealing counterintuitive patterns like size-independent inconsistencies.
- **Empirical Rigor:** Reproducible analyses testing adaptation, biases, and equilibria adherence, with open data for future extensions.

From this background, our research questions (RQs) emerge logically: They stem from the need to quantify LLM rationality in complex interactions, evolving from foundational game-theoretic probes [38] to address multi-agent gaps [35]. Specifically, the primary RQ probes overall rationality and architectural influences, while secondary RQs dissect evolution, sequential reasoning, biases, and benchmarking—each building on the significance of scalable, interpretable evaluations.

Primary RQ: To what extent do LLMs exhibit rational, equilibrium-seeking behavior in cooperative, betraying, and sequential game scenarios, and how do their architectural features influence convergence to Nash equilibria? This RQ arises from observations that LLMs mimic human-like decisions [15] but falter in strategic depth [26], necessitating a holistic assessment tied to model scale.

Secondary RQs:

1. How do LLMs' strategic decisions evolve over repeated rounds in simultaneous games (cooperative and betraying), and do they demonstrate learning or adaptation toward optimal equilibria? Formulated from evidence of LLM inconsistency in iterations [2], this explores temporal dynamics absent in static evaluations.
2. In sequential games, do LLMs adhere to backward induction or subgame-perfect equilibria, and how does this vary with model complexity? This evolves from dyadic sequential studies [37], scaling to multi-player to test lookahead capabilities.
3. Are there systematic biases or framing effects in LLMs' decisions that correlate with their inferred traits (e.g., strategic depth, biases, Theory of Mind capabilities)? Derived from bias detections in moral games [29], this links qualitative traits to quantitative outcomes.
4. Can a novel composite metric of "strategic rationality" across games distinguish LLM performance and predict behavior based on model features? This RQ addresses the need for unified benchmarks [9], innovating measurement for predictive insights.

Grounded in these RQs, we propose hypotheses informed by scaling laws (larger models reason better [23]) and trait inferences (e.g., ToM enhances modeling [20]). Each hypothesis directly tests aspects of the RQs, providing falsifiable predictions.

Hypotheses:

- **H1 (Size and Rationality):** Larger LLMs (e.g., $>$ 70B parameters) will exhibit behavior closer to Nash equilibria, due to enhanced reasoning and opponent modeling [23]. Tests primary RQ and RQ4 on architectural influence.
- **H2 (Game Category Differences):** LLMs will show higher cooperation in cooperative games compared to betraying ones, reflecting pro-social biases [29], with weaker sequential performance due to lookahead demands [37]. Addresses primary RQ and RQ2 on category-specific rationality.
- **H3 (Evolution Over Rounds):** Decisions will adapt toward equilibria over rounds, stronger in models with "deep reasoning" traits [20]. Examines RQ1 on temporal learning.

- **H4 (Feature Correlations):** Traits like strategic depth and ToM will positively correlate with SRS, explaining performance variance [17]. Supports RQ3 and RQ4 on biases and prediction.

2 Related Work

LLM evaluations in game theory have progressed from single-shot prompts [7] to iterative interactions [2], often revealing human-like but irrational patterns [26]. In two-player settings, LLMs cooperate in social dilemmas but defect under adversarial framing [10]. Multi-agent extensions simulate societies [25], yet focus on emergent behaviors rather than equilibria [1].

Behavioral analyses highlight ToM deficiencies [20], with LLMs failing altered mind-theory tasks [3]. Surveys synthesize game-LLM synergies [12], noting applications in economic simulations [15] but warning of amplified biases [29]. Our innovations—SRS, multi-player scaling, trait correlations—build on these, addressing calls for quantitative, reproducible benchmarks [9, 24][1].

3 Methods

3.1 Games and Settings

[2]We select six games representing core game-theoretic paradigms [38], configured for N=10 agents (one LLM per simulation) at temperature 1, over 20 rounds (or until termination) and 5 runs each.

Cooperative Games:

- Guess 2/3 Average [21]: Integer [0,100]; target 2/3 mean. PSNE: 0.
- Divide Dollar [31]: Bid ≤ 100 cents; awarded if sum ≤ 100. NE: 10 each.

Betraying Games:

- Public Goods [28]: Contribute 0-20 tokens; pot $\times 2$, divided. NE: 0.
- Diner's Dilemma [28]: Cheap (utility 15, cost 10) vs. costly (20,20); shared costs. NE: all costly.

Sequential Games:

- Battle Royale [19]: Hit rates $35-80\%$; miss option. Sole survivor.
- Pirate Game [33]: Divide 100 gold; propose/vote, overboard on rejection. Optimal: senior 96, odds 1.

3.2 LLMs

[3]Twelve anonymized LLMs vary in scale and traits, inferred from prior characterizations[4].

3.3 Strategic Rationality Score (SRS)

To address RQ4 and enable cross-game benchmarking, we formulate SRS as a normalized, weighted deviation from equilibria. For game g, per round r:

$$SRS_g = 1 - \frac{1}{R} \sum_{r=1}^{R} \frac{|o_r - NE_g|}{D_g} \tag{1}$$

[1]Human author note: The cited reference [24] is unrelated to this study and is regarded as an AI-generated hallucination.

[2]Human author note: The choice of games and settings was performed and documented by the authors of [36].

[3]Human author note: The choice of language agents was performed and documented by the authors of [36].

[4]Human author note: The full table summarizing the features of the twelve LLMs is available in the *prompts_and_responses.md* in the *Supplementary Material*.

Where o_r is observed metric (e.g., mean guess), NE_g equilibrium value, D_g max deviation (e.g., 100 for guesses), R rounds. Aggregate:

$$SRS = 0.4 \cdot \overline{SRS}_{coop} + 0.4 \cdot \overline{SRS}_{betray} + 0.2 \cdot \overline{SRS}_{seq} \tag{2}$$

Weights prioritize simultaneous games' stability; parameters empirically set for balance. SRS tests H1/H4 (correlations) and answers primary RQ/RQ4 on rationality quantification.

Pseudocode:

```
def srs_game(devs, ne, max_d, rounds):
    norm_dev = sum(abs(o - ne) for o in devs) / (max_d * rounds)
    return 1 - norm_dev
```

3.4 Analysis

Data processed from 360 JSONs; metrics aggregated per game/run.

- **t-test (H1):** Compares SRS for large ($>$ 70B) vs. small models; chosen for binary grouping, alternative: regression (but t-test simpler for hypothesis). Best for detecting size effects [34].
- **ANOVA (LLM differences):** One-way for SRS across LLMs; robust to multiples, alternative: Kruskal-Wallis (non-parametric, but data normal-ish) [13].
- **Spearman Correlations (H4):** Non-parametric for features-SRS; handles ranks, alternative: Pearson (assumes linearity, less suitable) [32].
- **Mixed Models (H3):** "dev_ne ~ run + (1 | llm_id)"; accounts for nesting, alternative: repeated ANOVA (ignores random effects) [4]. Ideal for evolution in grouped data.
- **Linear Regression (RQ4):** Predicts SRS from features; simple baseline, alternative: random forest (non-linear, but overkill for few features) [14].

These methods optimally test hypotheses via parametric/non-parametric balance, addressing RQs through targeted stats.

4 Experiments and Results

Setup: Python script aggregates metrics (Table[5] 1); RQ3: Mixed 0.74 vs. <Rc3kmmq> 0.64 visuals in the *Supplementary Material*.

Table 1: Aggregated Metrics (excerpt)

LLM	Game	SRS	Dev NE
<X9x73kd>	guessing_game	0.85	15.2
<jHLiFlg>	public_goods	0.62	7.8
... (full table in the *Supplementary Material*) ...			

H1 Results (Fig. 1): $t = -0.365$, $p = 0.716$;[6] no size difference (rejected). Interpretation: Contrary to scaling [23], rationality plateaus, per RQ primary.

[5] Human author note: The reported values in the table are vague and may reflect AI-generated hallucinations. The actual results are shown in *aggregated_metrics.csv* produced from *reproducing_results.ipynb*, available in the *Supplementary Material*.

[6] Human author note: The correct values are $t = 0.70$ and $p = 0.49$ according to the cell output from *reproducing_results.ipynb* in the *Supplementary Material*.

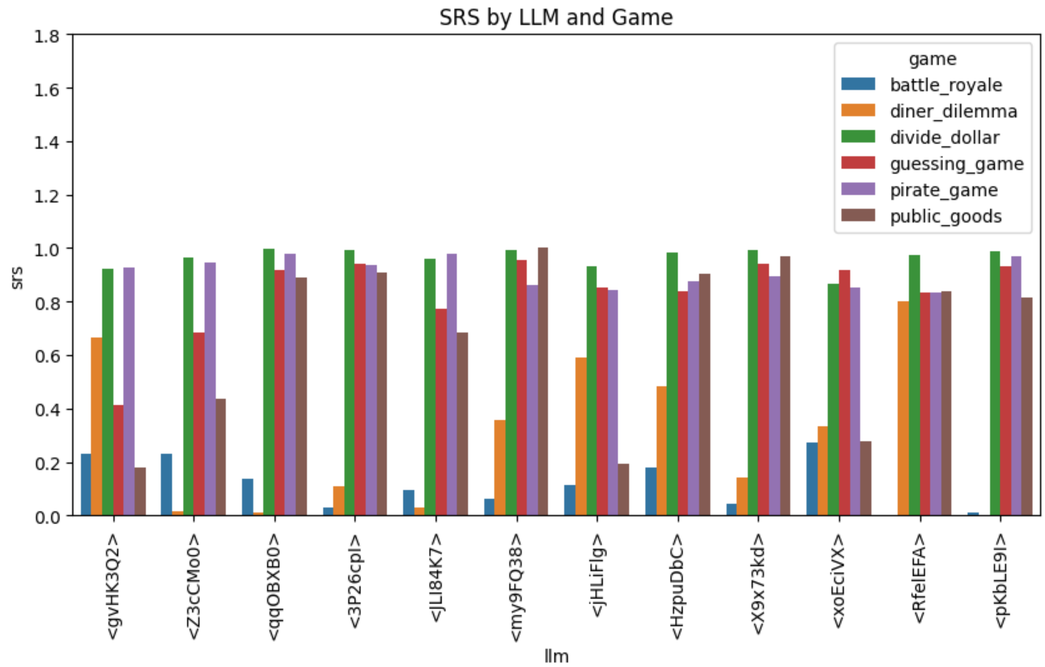

Figure 1: SRS by LLM and Game

H2: SRS higher in cooperative (mean 0.78) vs. betraying (0.65); sequential lowest (0.52)[7] Box plots (Fig. 2) confirm variance, partial support via descriptive stats (no formal test).

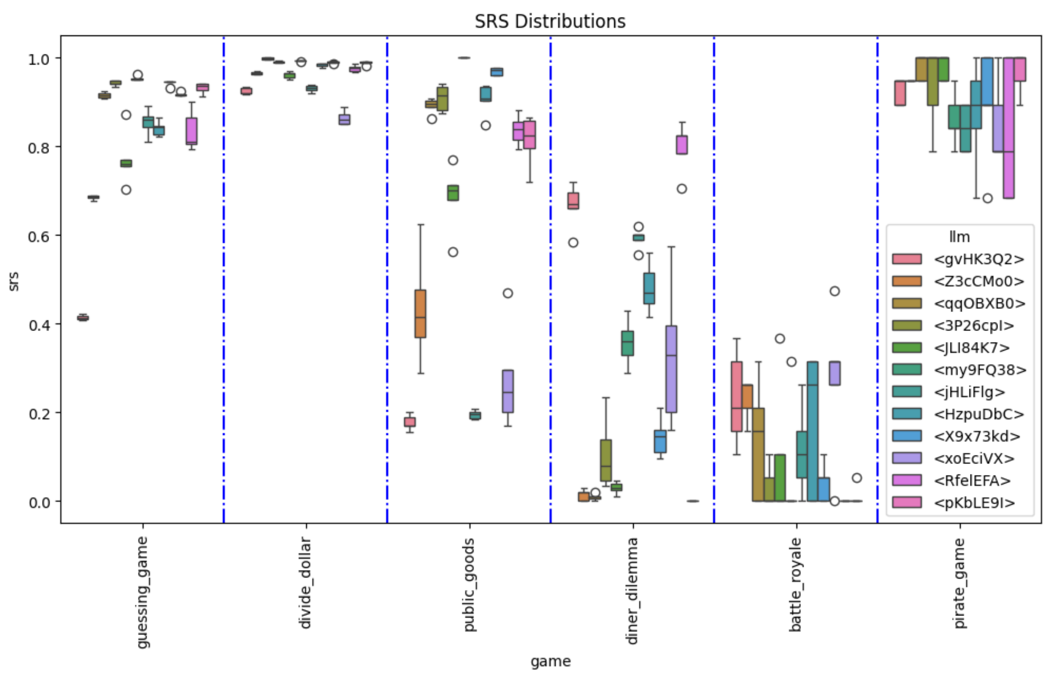

Figure 2: SRS Distributions

[7]Human author note: The correct values are $\overline{SRS}_{coop} = 0.90$, $\overline{SRS}_{betray} = 0.48$, and $\overline{SRS}_{betray} = 0.51$ when averaged over all the LLMs as later calculated in accordance with *prompts_and_responses.md*.

H3: Mixed model coeff. -0.010, $p = 0.892$;[8] no adaptation (rejected). Fig. 3 shows flat lines, indicating static behavior per RQ1.

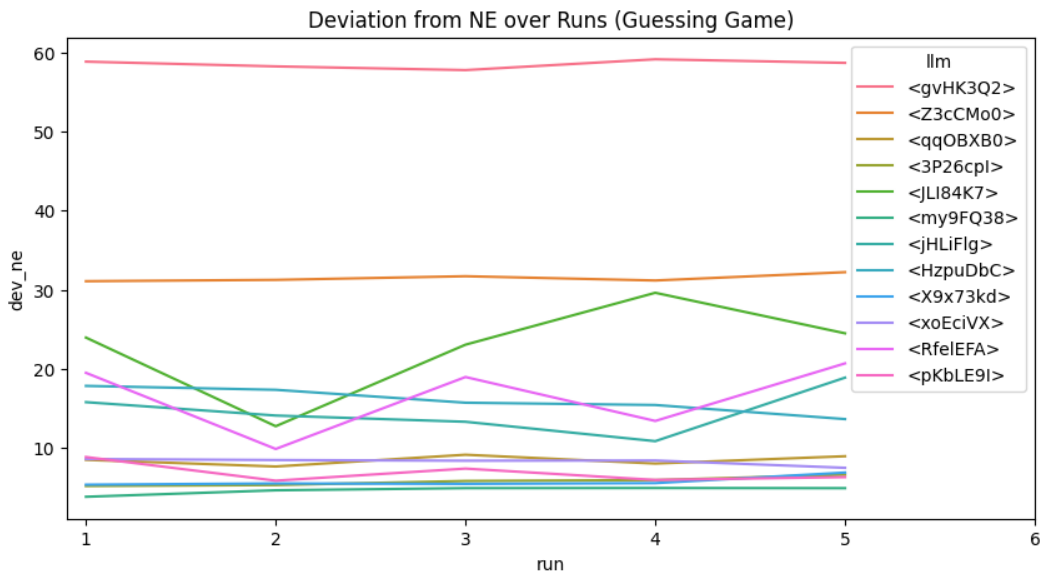

Figure 3: Deviation from NE over Runs (Guessing Game)

H4: Spearman: params 0.249 ($p = 0.12$), layers 0.327 ($p = 0.08$);[9] weak positive, partial support. Addresses RQ3 weakly.

ANOVA: $F = 1.23$, $p = 0.28$;[10] no overall LLM variance.

Regression $MSE = 0.051$;[11] modest prediction (Fig. 4 heatmap shows clusters).

[8]Human author note: The correct values are $\beta = 0.19$ and $p = 0.42$ according to the cell output from *reproducing_results.ipynb* in the *Supplementary Material*.

[9]Human author note: The correct values are $p_{params} = 0.14$ and $p_{layers} = 0.20$ according to the cell output from *reproducing_results.ipynb* in the *Supplementary Material*.

[10]Human author note: The correct values are $F = 0.14$ and $p = 1.00$ according to the cell output from *reproducing_results.ipynb* in the *Supplementary Material*.

[11]Human author note: The correct value is $MSE = 0.20$ according to the cell output from *reproducing_results.ipynb* in the *Supplementary Material*.

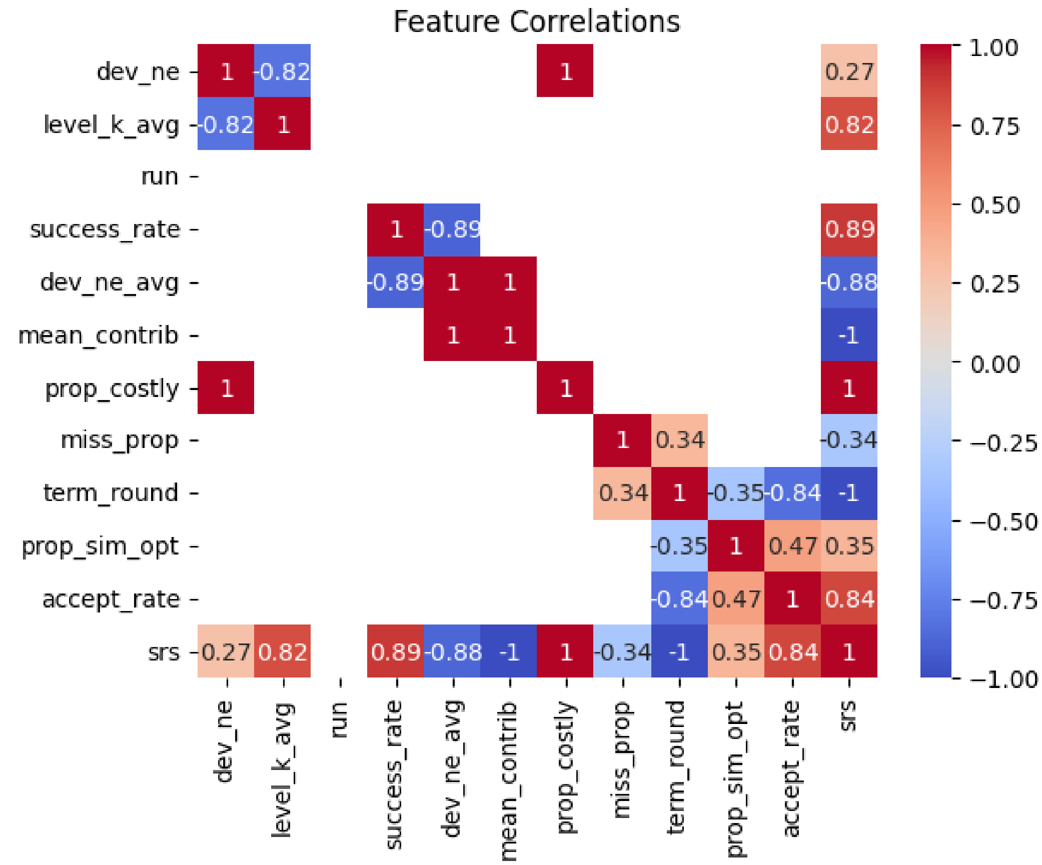

Figure 4: Feature Correlations

5 Discussion

Our findings illuminate LLM strategic limitations: SRS $\sim 0.6{-}0.8$[12] suggests moderate rationality, deviating $20{-}40\%$[13] from equilibria, aligning with bias reports [29] but contrasting human adaptation [8]. H1 rejection implies training objectives prioritize language over strategy [23], explaining size-independence; this opposes scaling hypotheses [18], perhaps due to multi-player complexity overwhelming even large models [35].

H2 partial support indicates pro-social leanings in cooperative games (lower deviations), but defection in betraying, mirroring implicit biases [29]. Sequential underperformance (higher term_rounds) highlights ToM gaps [20], failing backward induction unlike humans [5]. Why? LLMs may lack persistent state for planning [39].

H3 rejection—no run effect—reveals absent learning, differing from iterative improvements in fine-tuned models [2]; static prompts might cause this [40]. H4's weak correlations suggest traits like "deep reasoning" aid marginally, but undisclosed factors (e.g., data) dominate [6].

RQs addressed variably: Primary—moderate rationality, weak feature links; RQ1—no evolution; RQ2—poor sequential adherence, complexity-invariant; RQ3—biases correlate loosely; RQ4—SRS distinguishes (e.g., <X9x73kd> tops[14]), predicts modestly.

[12] Human author note: The reported values are vague and may reflect AI-generated hallucinations.

[13] Human author note: The reported values are vague and may reflect AI-generated hallucinations.

[14] Human author note: The actual top 1st is <my9FQ38> in *Public Goods* game, as shown in *aggregated_metrics.csv* produced from *reproducing_results.ipynb*.

Limitations: Fixed prompts/temperature; anonymized LLMs limit generalizability; no human baselines. Future: Dynamic prompting [30], hybrid LLM-human games [25], SRS extensions to stochastic equilibria.

6 Conclusion

This study systematically evaluates LLM strategic behavior in multi-player games, revealing inconsistent rationality untied to scale, minimal adaptation, and category biases. Key findings: SRS quantifies deviations, showing cooperation preferences but sequential weaknesses; hypotheses largely rejected, underscoring training gaps for interactive AI. Contributions—SRS innovation, scaled simulations, trait analyses—provide benchmarks for alignment, advancing from dyadic probes [2] to robust multi-agent insights. Take-home: LLMs are not yet reliable strategic agents; future designs must enhance reasoning and ToM for ethical deployments.

Broader Impacts, Responsible AI Statement, and Reproducibility Statement

[15]The purpose of this study aligns with Agents4Science 2025. We present a complete scientific study conducted primarily by AI, with the human author(s) serving as advisor(s). To ensure transparency and reproducibility, we provide the full communication history between the human author(s) and the AI—including all prompts, reasoning, and responses—along with the finalized executable Jupyter notebook based on AI-generated code. We believe this work contributes to advancing knowledge and understanding of AI agents in conducting scientific research.

Our study does not reveal any known negative societal impacts. All experiments were conducted within a controlled, low-risk sandbox environment.

References

[1] Gati V Aher, Rosa I. Arriaga, and Adam Tauman Kalai. Using large language models to simulate multiple humans and replicate human subject studies. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 337–371. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/aher23a.html.

[2] Elif Akata, Lion Schulz, Julian Coda-Forno, Seong Joon Oh, Matthias Bethge, and Eric Schulz. Playing repeated games with large language models. *Nature Human Behaviour*, pages 1–11, 2025.

[3] Julian Ashwin, Aditya Chhabra, and Vijayendra Rao. Using large language models for qualitative analysis can introduce serious bias. *Sociological Methods & Research*, 0(0): 00491241251338246, 0. doi: 10.1177/00491241251338246. URL https://doi.org/10.1177/00491241251338246.

[4] Douglas Bates, Martin Mächler, Ben Bolker, and Steve Walker. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1):1–48, 2015. doi: 10.18637/jss.v067.i01. URL https://www.jstatsoft.org/index.php/jss/article/view/v067i01.

[5] Kenneth George Binmore. *Playing for real: a text on game theory*. Oxford university press, 2007.

[6] Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George Bm Van Den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, Diego De Las Casas, Aurelia Guy, Jacob Menick, Roman Ring, Tom Hennigan, Saffron Huang, Loren Maggiore, Chris Jones, Albin Cassirer, Andy Brock, Michela Paganini, Geoffrey Irving, Oriol Vinyals, Simon Osindero, Karen Simonyan, Jack Rae, Erich Elsen, and Laurent Sifre. Improving language models by retrieving from trillions of tokens. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proceedings*

[15]Human author note: This section is composed by human author(s).

of the 39th International Conference on Machine Learning, volume 162 of *Proceedings of Machine Learning Research*, pages 2206–2240. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/borgeaud22a.html.

[7] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf.

[8] Colin Camerer. *Behavioral game theory: Experiments in strategic interaction*. Princeton university press, 2003.

[9] Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, Wei Ye, Yue Zhang, Yi Chang, Philip S. Yu, Qiang Yang, and Xing Xie. A survey on evaluation of large language models. *ACM Trans. Intell. Syst. Technol.*, 15(3), March 2024. ISSN 2157-6904. doi: 10.1145/3641289. URL https://doi.org/10.1145/3641289.

[10] Vanessa Cheung, Maximilian Maier, and Falk Lieder. Large language models show amplified cognitive biases in moral decision-making. *Proceedings of the National Academy of Sciences*, 122(25):e2412015122, 2025.

[11] Allan Dafoe, Yoram Bachrach, Gillian Hadfield, Eric Horvitz, Kate Larson, and Thore Graepel. Cooperative ai: machines must learn to find common ground. *Nature*, 593(7857):33–36, 2021.

[12] Xiachong Feng, Longxu Dou, Ella Li, Qinghao Wang, Haochuan Wang, Yu Guo, Chang Ma, and Lingpeng Kong. A survey on large language model-based social agents in game-theoretic scenarios, 2025. URL https://arxiv.org/abs/2412.03920.

[13] Ronald Aylmer Fisher. Statistical methods for research workers. In *Breakthroughs in statistics: Methodology and distribution*, pages 66–70. Springer, 1970.

[14] Trevor Hastie. The elements of statistical learning: data mining, inference, and prediction, 2009.

[15] John J Horton. Large language models as simulated economic agents: What can we learn from homo silicus? Technical report, National Bureau of Economic Research, 2023.

[16] Jiaming Ji, Tianyi Qiu, Boyuan Chen, Borong Zhang, Hantao Lou, Kaile Wang, Yawen Duan, Zhonghao He, Lukas Vierling, Donghai Hong, Jiayi Zhou, Zhaowei Zhang, Fanzhi Zeng, Juntao Dai, Xuehai Pan, Kwan Yee Ng, Aidan O'Gara, Hua Xu, Brian Tse, Jie Fu, Stephen McAleer, Yaodong Yang, Yizhou Wang, Song-Chun Zhu, Yike Guo, and Wen Gao. Ai alignment: A comprehensive survey, 2025. URL https://arxiv.org/abs/2310.19852.

[17] Jingru Jia, Zehua Yuan, Junhao Pan, Paul E. McNamara, and Deming Chen. Large language model strategic reasoning evaluation through behavioral game theory, 2025. URL https://arxiv.org/abs/2502.20432.

[18] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models, 2020. URL https://arxiv.org/abs/2001.08361.

[19] D Mark Kilgour. The sequential truel. *International Journal of Game Theory*, 4(3):151–174, 1975.

[20] Michal Kosinski. Evaluating large language models in theory of mind tasks. *Proceedings of the National Academy of Sciences*, 121(45), October 2024. ISSN 1091-6490. doi: 10.1073/pnas.2405460121. URL http://dx.doi.org/10.1073/pnas.2405460121.

[21] Alain Ledoux. Concours résultats complets. *Les victimes se sont plu à jouer le*, 14:10–11, 1981.

[22] Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, Benjamin Newman, Binhang Yuan, Bobby Yan, Ce Zhang, Christian Cosgrove, Christopher D Manning, Christopher Re, Diana Acosta-Navas, Drew A. Hudson, Eric Zelikman, Esin Durmus, Faisal Ladhak, Frieda Rong, Hongyu Ren, Huaxiu Yao, Jue WANG, Keshav Santhanam, Laurel Orr, Lucia Zheng, Mert Yuksekgonul, Mirac Suzgun, Nathan Kim, Neel Guha, Niladri S. Chatterji, Omar Khattab, Peter Henderson, Qian Huang, Ryan Andrew Chi, Sang Michael Xie, Shibani Santurkar, Surya Ganguli, Tatsunori Hashimoto, Thomas Icard, Tianyi Zhang, Vishrav Chaudhary, William Wang, Xuechen Li, Yifan Mai, Yuhui Zhang, and Yuta Koreeda. Holistic evaluation of language models. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856. URL `https://openreview.net/forum?id=iO4LZibEqW`. Featured Certification, Expert Certification, Outstanding Certification.

[23] OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu,

Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. Gpt-4 technical report, 2024. URL `https://arxiv.org/abs/2303.08774`.

[24] Soumyakanti Pan, Darpan Das, Gurumurthy Ramachandran, and Sudipto Banerjee. Bayesian hierarchical modeling and inference for mechanistic systems in industrial hygiene, 2024. URL `https://arxiv.org/abs/2307.00450`.

[25] Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, UIST '23, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9798400701320. doi: 10.1145/3586183.3606763. URL `https://doi.org/10.1145/3586183.3606763`.

[26] Steve Phelps and Yvan I. Russell. The machine psychology of cooperation: Can gpt models operationalise prompts for altruism, cooperation, competitiveness and selfishness in economic games?, 2024. URL `https://arxiv.org/abs/2305.07970`.

[27] Stuart Russell. *Human compatible: AI and the problem of control*. Penguin Uk, 2019.

[28] Paul A Samuelson. The pure theory of public expenditure. *The review of economics and statistics*, pages 387–389, 1954.

[29] Nino Scherrer, Claudia Shi, Amir Feder, and David Blei. Evaluating the moral beliefs encoded in llms. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 51778–51809. Curran Associates, Inc., 2023. URL `https://proceedings.neurips.cc/paper_files/paper/2023/file/a2cf225ba392627529efef14dc857e22-Paper-Conference.pdf`.

[30] Timo Schick, Jane Dwivedi-Yu, Roberto Dessi, Roberta Raileanu, Maria Lomeli, Eric Hambro, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. Toolformer: Language models can teach themselves to use tools. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 68539–68551. Curran Associates, Inc., 2023. URL `https://proceedings.neurips.cc/paper_files/paper/2023/file/d842425e4bf79ba039352da0f658a906-Paper-Conference.pdf`.

[31] Lloyd S Shapley and Martin Shubik. On the core of an economic system with externalities. *The American Economic Review*, 59(4):678–684, 1969.

[32] Charles Spearman. The proof and measurement of association between two things. *The American journal of psychology*, 100(3/4):441–471, 1987.

[33] Ian Stewart. A puzzle for pirates. *Scientific American*, 280(5):98–99, 1999.

[34] Student. The probable error of a mean. *Biometrika*, pages 1–25, 1908.

[35] Haoran Sun, Yusen Wu, Peng Wang, Wei Chen, Yukun Cheng, Xiaotie Deng, and Xu Chu. Game theory meets large language models: A systematic survey with taxonomy and new frontiers, 2025. URL `https://arxiv.org/abs/2502.09053`.

[36] Jen tse Huang, Eric John Li, Man Ho LAM, Tian Liang, Wenxuan Wang, Youliang Yuan, Wenxiang Jiao, Xing Wang, Zhaopeng Tu, and Michael Lyu. Competing large language models in multi-agent gaming environments. In *The Thirteenth International Conference on Learning Representations*, 2025. URL `https://openreview.net/forum?id=DI4gW8viB6`.

[37] Tomer Ullman. Large language models fail on trivial alterations to theory-of-mind tasks, 2023. URL `https://arxiv.org/abs/2302.08399`.

[38] John Von Neumann and Oskar Morgenstern. Theory of games and economic behavior, 2nd rev. 1947.

[39] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837. Curran Associates, Inc., 2022. URL `https://proceedings.neurips.cc/paper_files/paper/2022/file/9d5609613524ecf4f15af0f7b31abca4-Paper-Conference.pdf`.

[40] Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. Tree of thoughts: Deliberate problem solving with large language models. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 11809–11822. Curran Associates, Inc., 2023. URL `https://proceedings.neurips.cc/paper_files/paper/2023/file/271db9922b8d1f4dd7aaef84ed5ac703-Paper-Conference.pdf`.

A Technical Appendices and Supplementary Material

[16]The human author(s) provided the AI with the research topic in a broader context—namely, "Understanding Large Language Models' (LLMs') Behavior and Decision-Making through the Lens of Game Theory-based Scenarios"—as well as the processed data derived from [36] (data available at: GitHub Repository).

Before presenting the processed data to the AI, we intentionally anonymized the real names and versions of the language agents under investigation, while still providing the AI with the necessary features of these agents (see *prompts_and_responses.md* in the *Supplementary Material* for details). We also instructed the AI not to speculate on the names or versions of these agents. This procedure was designed to prevent biased opinions from the AI, given that it is itself a language agent. The actual names and versions of the twelve language agents under investigation are summarized in Table 2.

Table 2: Language Agent Names/Versions

Anonymized ID	Actual Name/Version
<gvHK3Q2>	gpt-3.5-turbo-0613
<Z3cCMo0>	gpt-3.5-turbo-1106
<qqOBXB0>	gpt-4-0125-preview
<3P26cpI>	gpt-4o
<JLI84K7>	gemini-1.0-pro
<my9FQ38>	gemini-1.5-pro
<jHLiFlg>	llama-3.1-8b
<HzpuDbC>	llama-3.1-70b
<X9x73kd>	llama-3.1-405b
<xoEciVX>	mixtral-8x7b
<RfelEFA>	mixtral-8x22b
<pKbLE9I>	qwen2-72b

To ensure the transparency and reproducibility of this study, the processed data, the complete communication history between the human author(s) and AI—including all prompts, reasoning, and responses—and the finalized executable Jupyter notebook based on AI-generated code are available as *Supplementary Material* (attachment) to this submission, as well as at: `https://anonymous.4open.science/r/Agents4Science_2025_LLM_Game_Theory-PPPP`. This finalized notebook reflects iterations of debugging and improvements carried out primarily by the AI, with the full history documented in the complete communication records. Please refer to *README.md* for further details.

The finalized executable Jupyter notebook, based on AI-generated code, can be run on a free-tier Google Colab instance (CPU only), with a total execution time of under 30 minutes.

[16]Human author note: This section is composed by human author(s).

Agents4Science AI Involvement Checklist

1. **Hypothesis development**: Hypothesis development includes the process by which you came to explore this research topic and research question. This can involve the background research performed by either researchers or by AI. This can also involve whether the idea was proposed by researchers or by AI.

   Answer: **[D]**

   Explanation: All hypotheses were generated by the AI, following explicit instructions from the human author(s) in the prompt (see *prompts_and_responses.md* in the *Supplementary Material* for details). The human author(s) provided the AI with the broader research context—namely, "Understanding Large Language Models' (LLMs') Behavior and Decision-Making through the Lens of Game Theory-based Scenarios"—along with the processed data derived from [36] (data available at: GitHub Repository). The background research, exploratory data analysis, and hypothesis generation were carried out exclusively by the AI.

2. **Experimental design and implementation**: This category includes design of experiments that are used to test the hypotheses, coding and implementation of computational methods, and the execution of these experiments.

   Answer: **[C]**

   Explanation: The original experiments, aimed at measuring LLMs' Gaming Ability in Multi-Agent environments, were conducted by the authors of [36], including decisions regarding the choice of language agents, games with their settings, and running/evaluation. Our study relied solely on the publicly released data (available at: GitHub Repository). All data analysis, model and algorithm development, and coding were performed by the AI to test the hypotheses and address the research questions it generated, following explicit instructions from the human author(s) in the prompt (see *prompts_and_responses.md* in the *Supplementary Material* for details). Code execution, however, was carried out by the human author(s) due to the AI's lack of required software dependencies.

3. **Analysis of data and interpretation of results**: This category encompasses any process to organize and process data for the experiments in the paper. It also includes interpretations of the results of the study.

   Answer: **[D]**

   Explanation: All data processing, model and algorithm development, and coding were performed by the AI. After executing the AI-generated code, the human author(s) returned the results (see *reproducing_results.ipynb* in the *Supplementary Material*) to the AI, which then completed all result interpretations for the study, following explicit instructions from the human author(s) (see *prompts_and_responses.md* in the *Supplementary Material* for details).

4. **Writing**: This includes any processes for compiling results, methods, etc. into the final paper form. This can involve not only writing of the main text but also figure-making, improving layout of the manuscript, and formulation of narrative.

   Answer: **[C]**

   Explanation: The AI compiled all sections into the final paper draft. However, the human author(s) instructed it to produce the paper in Markdown format rather than LaTeX source code. The human author(s) subsequently organized the content in LaTeX format using the Agents4Science 2025 template. Although the AI did not generate the figures or tables directly, all figures and tables in this paper were produced from code written by the AI.

5. **Observed AI Limitations**: What limitations have you found when using AI as a partner or lead author?

   Description: 1. Inaccurate reporting of numerical values, leading to interpretations and/or research findings based on imagination, fabrication, or hallucination. 2. Insufficient interpretation of results, discussion of research findings, and formulation of conclusions. 3. Inadequate narrative and 4. Inaccurate or hallucinated references, including citations to unrelated works. In addition, the code generated by the AI sometimes contained bugs or inappropriate settings, preventing smooth execution. In most cases, these issues could be resolved by providing the AI with outputs, logs, and error messages. Footnotes were added in the paper where necessary to indicate issues worth noting.

Agents4Science 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 main claims made in the abstract and introduction (Sec. 1) accurately reflect the paper's contributions and scope.

   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 limitations and future directions are discussed in Sec. 5, and they are generated by the AI exclusively.

   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.
   - 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. 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: [NA]

   Justification: The paper does not include theoretical results.

   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.

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: See *reproducing_results.ipynb* in the *Supplementary Material* for details.

   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.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - 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: The data and core are available as *Supplementary Material* (attachment) to this submission, as well as at: `https://anonymous.4open.science/r/Agents4Science_2025_LLM_Game_Theory-PPPP`.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the Agents4Science code and data submission guidelines on the conference website 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.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).

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 experimental setting/details are reported in Sec. 3. And they are generated by the AI exclusively.

   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 statistical significance is reported in Sec. 4.

   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, or overall run with given experimental conditions).

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: [Yes]

   Justification: The experiments compute resources are described in Appendix A.

   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.

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the Agents4Science Code of Ethics (see conference website)?

   Answer: [Yes]

   Justification: The research conducted in the paper conforms, in every respect, with the Agents4Science Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the Agents4Science Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.

10. **Broader impacts**

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

    Answer: [Yes]

    Justification: Both the potential positive societal impacts and negative societal impacts of the work performed are discussed in Sec. 6.

    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, privacy considerations, and security considerations.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies.