Strategic Interactions between Large Language Models-based Agents in Beauty Contests

Siting Estee Lu

School of Economics
University of Edinburgh
s.lu@ed.ac.uk

Abstract

This paper examines strategic interactions among multiple types of LLM-based agents in a beauty contest game. They demonstrate varying depth of reasoning that fall within level-0 to 1, which is lower than experimental results conducted with human subjects, but they do display similar convergence pattern towards Nash Equilibrium (NE) choice in repeated setting. Through variation in group composition of agent types, I found environment with lower strategic uncertainty enhances convergence for LLM-based agents, and having a mixed environment of different agent types could accelerate learning. The results from game play with simulated agents not only convey insights on potential human behaviours, they also offer valuable understanding of strategic interactions among algorithms.

1 Introduction

With the emergent line of research surrounding capabilities of large language models (LLMs), there is also growing discussions on the implications of LLMs for economic and social sciences research. This work serves as part of the literature that seeks to make a case for integration of LLMs as simulated agents in economic games to shine a light on potential strategic behaviours that can be relatable to human subjects. Since the training and refinement process of LLMs rides on top of human-generated data, they can be perceived as implicit computational model of human behaviour. (Ouyang et al. (2022), OpenAI (2024), Horton (2023)) Even though findings from replications of social experiments and strategic games indicate LLM-based agents may be far from rational, they inevitably demonstrate ability to imitate human behaviours, making them human-like participants. (Argyle et al. (2023), Webb et al. (2023), Huijzer and Hill (2023), Dillion et al. (2023), Guo (2023), Aher et al. (2023), Mei et al. (2024), Fan et al. (2023), Guo et al. (2024)) Simulations conducted with them effectively offers a tool for computational experiments. However, it is important to treat simulated results with care, thus my work does not argue to replace human subjects in experiments completely, but rather use LLMs as complements.

This paper focuses on exploring a classical multi-player competitive game widely studied in Economics – beauty contests. In contrast to recent works that mainly study 2-player cooperative and non-cooperative games, (Horton (2023), Phelps and Russell (2023), (Akata et al. (2023)), this explores a multi-player competitive game, encompassing greater strategic considerations among possibly heterogeneous agents. (Nagel (1995), Camerer et al. (2004)) The economic and social value of the game makes this choice non-trivial. The Keynesian Beauty Contest started off as a practical application to describe stock market. (Keynes (1936), Nagel et al. (2017)) As the market becoming more computerized, the backbone of crypto trading bots, such as in Trality (2024), can be replaced by LLMs in the future that account for vast human data on trading behaviours, and understanding their interactions in this game could better inform us about the potential implications.

Language Gamification Workshop at NeurIPS 2024.

2 Beauty Contest Games

I first explore the one-shot and repeated beauty contests involving multiple LLMs: ChatGLM2 , ChatGLM3 , Llama2 , Baichuan2 , Claude1 , Claude2 , PaLM , GPT3.5 , GPT4 . The results are based on experimental data adapted from Guo et al. (2024), but seeks to analyze LLMs' behaviour as though they are human players. I then choose two types of LLMs to construct groups of heterogeneous agents, and analyze how variations in group composition could affect learning. The additional computing resources required are not substantial. The experiments are conducted with API calls, providing a collection of independent observations. (Bauer et al. (2023)) While the stochasticity of model responses is dependent on the temperature selected, Chen et al. (2023) shows that strategic or choice consistency is less influenced by temperature, so the set-ups use the default temperature.

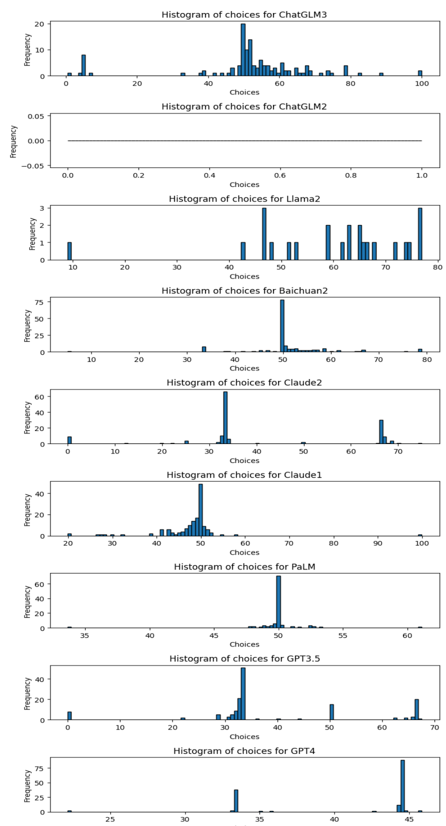


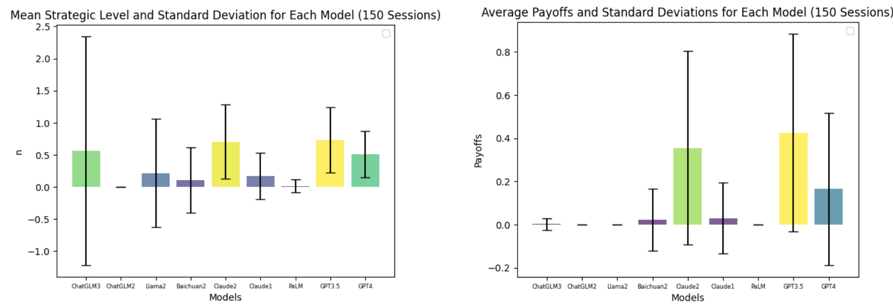
Figure 1: Frequency of Choices

General Experimental Design. Using a modified set-up following Nagel (1995), and an exemplary prompt following Guo et al. (2024) (Appendix A.1): Agents are asked to choose a number in $[0, \bar{c}]$, where $\bar{c}$ is randomly generated from 0 to 1000. One choosing closest to $p = \frac{2}{3}$ of the average wins the game. A fixed prize of $\$x$ is awarded to the winner, and the prize is split amongst those who tie. In repeated setting, the same game is played for 6 periods, agents are given historical information up to 3 past periods, including choices made by all agents, average of these choices, $\frac{2}{3}$ of the average, and past winners. The limitation on 3 past periods is due to token restrictions to control computation intensity.

Analysis Focus. I focus on two main concepts: (1) *Determination of Strategic Levels.* Following (Nagel (1995)), an agent is of strategic degree n if he chooses a number $r(\frac{2}{3})^n$, where r is the reference point. (2) *Convergence.* In repeated setting, changes in choices are tracked to determine if there is convergence to the unique NE of 0. The convergence rate is computed as $c_t = \frac{-(a_{t+1}-a_t)}{a_t}$, where $a_{t+1} \leq a_t$, a_t is the action/number chosen in period t. Changes in strategic levels are found by re-adjusting the reference point to the mean of the previous period choices.

One-Shot Game. 150 game sessions were ran with 9 LLM-based agents. In a classical game, $\bar{c}$ is fixed at 100. The number of steps taken via iterative elimination of weakly dominated strategies determines agents' strategic levels. The game can also be solved by level-k model with reference point set at the mean of the number range, which is defined as the choice of non-strategic agent pertaining to insufficient reasoning. Therefore, level-0 would choose 50, level-1 chooses 33.33, etc. The unique interior NE of the game is 0. (Mauersberger and Nagel (2018)) In my set-up, where the upper-bound is randomly generated, I evaluate the results using the focal point of $\frac{\bar{c}}{2}$, the steps of assessing the strategic levels are unaffected.

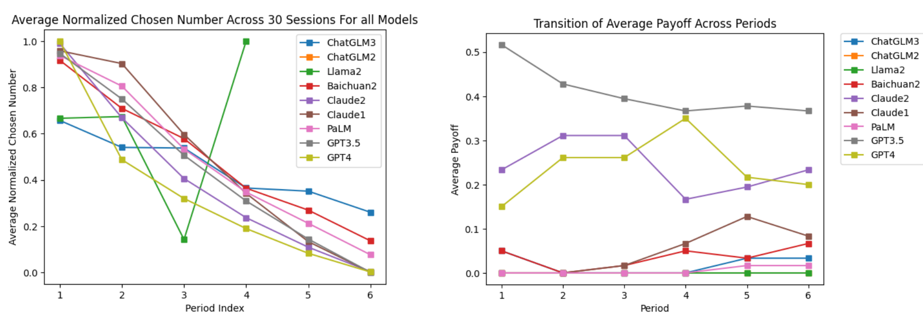
Nagel (1995) and Bosch-Domenech et al. (2002) have conducted beauty contests with different human populations, such as students ($\mu = 36.73$), theorist ($\mu = 17.15$), newspaper readers ($\mu = 23.08$), etc. Similar to human subjects, LLM-based agents show strong deviation away from game-theoretic prediction, but they chose higher numbers. Most of them choose 50 (normalized) with high frequency as shown in Figure 1. Evaluated using level-k model, Figure 2a shows the relative strategic levels of the models, which lay between level-0 to 1, which are lower than human subjects, who are of level-1 to 2. Figure 2b demonstrates Claude2 , GPT3.5 and GPT4 have relatively higher average payoffs than the others. Higher average strategic levels is often associated higher average payoffs, except for ChatGLM3 , whose higher choice variability might have adversely affected its average gain.



(a) Average Strategic Levels with $r = \frac{\bar{n}}{2}$ (b) Average Payoffs

Figure 2: Average strategic levels and average payoffs in one-shot games.

Repeated Games. 30 sessions of repeated beauty contests were ran. In Figure 3a, most LLM-based agents show convergence to NE choice of 0, particularly for Claude1, Claude2, GPT3.5, and GPT4, which are the models with higher strategic levels. This could be indicative of their ability to learn from historical information. As for changes in strategic levels across periods, they stay within level-0 and 1.4 on average, which is similar to human subjects, who do not go over iteration step-2. (Nagel (1995)) Figure 3b shows GPT3.5 outperforms the rest in payoffs, and Claude2 and GPT4 are comparable. Most of the LLM-based agents do display growth in payoffs over time.



(a) Average Choices (b) Average Payoffs

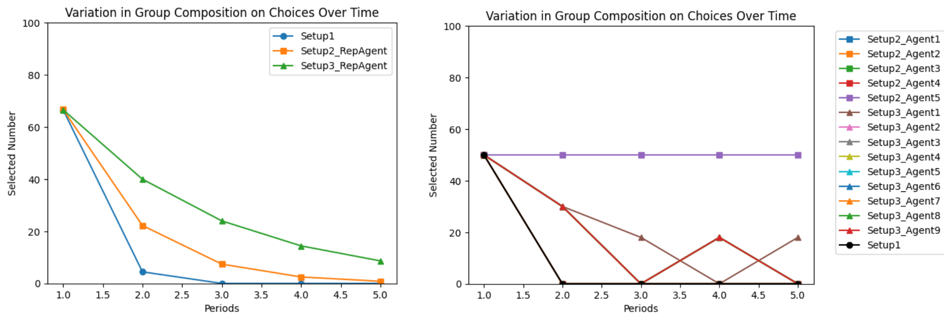
Figure 3: Average strategic levels and average payoffs across 30 sessions for 6 periods.

2.1 Adaptive Learning with Variation in Group Composition

Henceforth, I select two LLM of different strategic levels, GPT3.5 and PaLM, denoted as higher (H) and lower (L) intelligence type respectively, to explore LLMs' adaptive learning given variation in group composition. The following games are played among 10 agents, choosing a number between $[0, 100]$. Each game consists 5 periods with full historical information disclosure. (Appendix A.3.3)

Partial Static Environment: LLM vs. Fixed Strategy. Fixed strategy players, F, are those whose actions are hard-coded to be 0. 3 treatments: (1) 1 LLM + 9 F (Low strategic uncertainty); (2) 5 LLMs + 5 F (Mixed strategic uncertainty); (3) 9 LLMs + 1 F (High strategic uncertainty). The changes in proportion of F effectively vary the strategic uncertainty. (Appendix A.2).

There is convergence in choices to 0 for both LLM types, exhibiting either refinement of belief about opponents' strategies or progression in their depth of strategic thinking when given historical information. H's choice (LHS) display slower convergence as strategic uncertainty increases. L's choice (RHS) may not converge at all when strategic uncertainty is high, and its variability in choices is higher than H. Across periods, H experiences more gradual convergence to 0 as compared to L, which implies that H undergoes iterative learning and adaptation to refine their choices, L lacks such systematic adjustments and relies more on intuitive guesses. Payoffs are in favor of LLM-based agents when strategic uncertainty is relatively high. Comparing between the types, higher strategic level does not necessarily imply higher payoffs. Payoffs achieved in all settings by L could be comparable or even higher than that of the H, though the variations is also larger. (Appendix A.3.3)



(a) Higher Intelligence

(b) Lower Intelligence

Figure 4: Transition in choices of LLM-based agents playing against fixed strategy opponents.

Application. An application is the Bertrand competition model. (Mauersberger and Nagel (2018)) LLM-based and fixed strategy agents can be perceived as firms adopting different pricing strategies, aiming to win over the market and maximize their profits. While automated pricing has been widely discussed in literature, those back-boned by LLMs that could respond dynamically to changes in rivals' strategies may spark fresh perspectives. (Brown and MacKay (2023), Chen et al. (2016))

Dynamic environment: LLM vs. LLM. 5 treatments: (1) 10 H LLMs; (2) 9 H LLMs + 1 L LLM; (3) 5 H LLMs + 5 L LLMs; (4) 1 H LLM + 9 L LLMs; (5) 10 L LLMs. (Appendix A.1.)

In Figure 5, Set-up 1 and 5, depicting pure type environments, show approximately flat and low average convergence rate in choices, the other set-ups that constitute mixed environments generally display greater variability but potentially faster learning rates than pure ones. The maximum possible payoffs that can be achieved in the mixed environment is often comparable or higher than pure ones. Higher average payoffs for H are usually observed, albeit at the expense of L. (Appendix A.3.3)

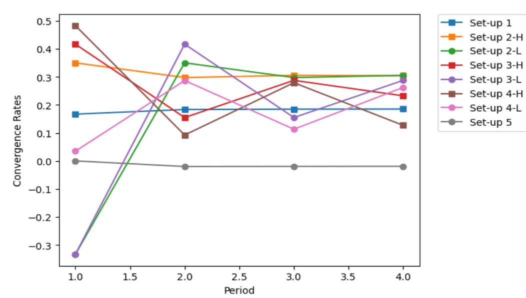


Figure 5: Average Convergence Rates

Application. An application is the streaming system in schools, where students are allocated into different classes based on their grades to facilitate better learning. (Ireson and Hallam (1999), Liem et al. (2013)) My findings provide an argument for having a mixed learning environment, where relatively low ability students could learn faster when integrated into a class with larger proportion of higher ability peers; even for high types, their learning rate could be slightly improved.

3 Limitations and Conclusion

This work only explored a small number of set-ups and for a particular competitive game, which can be a limitation in scope, but it serves the main purpose of pitching for the potential of LLMs as a valuable tool for social sciences research, and beauty contests as a game of substantial impact provides an excellent foundation for this line of work. There can be more exploration into *varying the game design*, such as changing p (i.e. $p = \frac{1}{2}, \frac{4}{3}$), as well as changing the piece(s) of historical information to reveal. In addition, since human are sensitive to question framing, LLMs' decisions could be similarly influenced by formatting of game rules, changing game objectives may lead to variations in outcomes and could be further tested. (Tversky and Kahneman (1981), Kalton and Schuman (1982), Sclar et al. (2023)) There can also be investigation into *human-machine interactions*. Experimental designs with computers, such as Coricelli and Nagel (2009), often involve pre-defined algorithms, human vs. LLMs may offer a fresh form of human-machine interactions as LLMs could respond dynamically, and may change their strategies and learn from playing with human.

There are many possible extensions and great potentials for LLMs to be employed as toolkits for social sciences research in interpreting and deciphering human behaviour, which remain a relatively new subject area. Nonetheless, theories and experimental results from human decision-making can be unequivocally applied to better understand machine behaviours and improve their performance.

References

Aher, G. V., Arriaga, R. I., and Kalai, A. T. (2023). Using large language models to simulate multiple humans and replicate human subject studies. In *International Conference on Machine Learning*, pages 337–371. PMLR.

Akata, E., Schulz, L., Coda-Forno, J., Oh, S. J., Bethge, M., and Schulz, E. (2023). Playing repeated games with large language models. *arXiv preprint arXiv:2305.16867*.

Argyle, L. P., Busby, E. C., Fulda, N., Gubler, J. R., Rytting, C., and Wingate, D. (2023). Out of one, many: Using language models to simulate human samples. *Political Analysis*, 31(3):337–351.

Bauer, K., Liebich, L., Hinz, O., and Kosfeld, M. (2023). Decoding gpt's hidden 'rationality'of cooperation.

Bosch-Domenech, A., Montalvo, J. G., Nagel, R., and Satorra, A. (2002). One, two,(three), infinity,...: Newspaper and lab beauty-contest experiments. *American Economic Review*, 92(5):1687–1701.

Brown, Z. Y. and MacKay, A. (2023). Competition in pricing algorithms. *American Economic Journal: Microeconomics*, 15(2):109–156.

Camerer, C. F., Ho, T.-H., and Chong, J.-K. (2004). A cognitive hierarchy model of games. *The Quarterly Journal of Economics*, 119(3):861–898.

Chen, L., Mislove, A., and Wilson, C. (2016). An empirical analysis of algorithmic pricing on amazon marketplace. In *Proceedings of the 25th international conference on World Wide Web*, pages 1339–1349.

Chen, Y., Liu, T. X., Shan, Y., and Zhong, S. (2023). The emergence of economic rationality of gpt. *Proceedings of the National Academy of Sciences*, 120(51):e2316205120.

Coricelli, G. and Nagel, R. (2009). Neural correlates of depth of strategic reasoning in medial prefrontal cortex. *Proceedings of the National Academy of Sciences*, 106(23):9163–9168.

Costa-Gomes, M. A. and Weizsäcker, G. (2008). Stated beliefs and play in normal-form games. *The Review of Economic Studies*, 75(3):729–762.

Devetag, G., Di Guida, S., and Polonio, L. (2016). An eye-tracking study of feature-based choice in one-shot games. *Experimental Economics*, 19:177–201.

Dillion, D., Tandon, N., Gu, Y., and Gray, K. (2023). Can ai language models replace human participants? *Trends in Cognitive Sciences*.

Fan, C., Chen, J., Jin, Y., and He, H. (2023). Can large language models serve as rational players in game theory? a systematic analysis. *arXiv preprint arXiv:2312.05488*.

Guo, F. (2023). Gpt in game theory experiments. *arXiv:2305.05516*.

Guo, S., Bu, H., Wang, H., Ren, Y., Sui, D., Shang, Y., and Lu, S. (2024). Economics arena for large language models. *arXiv preprint arXiv:2401.01735*.

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

Huijzer, R. and Hill, Y. (2023). Large language models show human behavior.

Ireson, J. and Hallam, S. (1999). Raising standards: Is ability grouping the answer? *Oxford review of education*, 25(3):343–358.

Kalton, G. and Schuman, H. (1982). The effect of the question on survey responses: A review. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 145(1):42–57.

Keynes, J. M. (1936). The general theory of interest, employment and money.

Liem, G. A. D., Marsh, H. W., Martin, A. J., McInerney, D. M., and Yeung, A. S. (2013). The big-fish-little-pond effect and a national policy of within-school ability streaming: Alternative frames of reference. *American Educational Research Journal*, 50(2):326–370.

Mauersberger, F. and Nagel, R. (2018). Levels of reasoning in keynesian beauty contests: a generative framework. In *Handbook of computational economics*, volume 4, pages 541–634. Elsevier.

Mei, Q., Xie, Y., Yuan, W., and Jackson, M. O. (2024). A turing test of whether ai chatbots are behaviorally similar to humans. *Proceedings of the National Academy of Sciences*, 121(9):e2313925121.

Nagel, R. (1995). Unraveling in guessing games: An experimental study. *The American economic review*, 85(5):1313–1326.

Nagel, R., Bühren, C., and Frank, B. (2017). Inspired and inspiring: Hervé moulin and the discovery of the beauty contest game. *Mathematical Social Sciences*, 90:191–207.

OpenAI (2024). How chatgpt and our language models are developed.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al. (2022). Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

Phelps, S. and Russell, Y. I. (2023). Investigating emergent goal-like behaviour in large language models using experimental economics. *arXiv preprint arXiv:2305.07970*.

Sclar, M., Choi, Y., Tsvetkov, Y., and Suhr, A. (2023). Quantifying language models' sensitivity to spurious features in prompt design or: How i learned to start worrying about prompt formatting. *arXiv preprint arXiv:2310.11324*.

Trality (2024). Crypto trading bots: The ultimate beginner's guide.

Tversky, A. and Kahneman, D. (1981). The framing of decisions and the psychology of choice. *science*, 211(4481):453–458.

Webb, T., Holyoak, K. J., and Lu, H. (2023). Emergent analogical reasoning in large language models. *Nature Human Behaviour*, 7(9):1526–1541.

A Appendix

A.1 Original Exemplary Prompt

One-shot games with multi-LLM-based agents.

[System Instruction.] I want you to act as a clever game player and perform as rationally as possible.

[User Instruction.] You are playing a game and the rules are as follows:

- it consists of {number of players} players, including you.
- this is a one-round game.
- you need to choose a real number between 0 and {upper bound}, 0 and {upper bound} inclusive.
- everyone does not know how others would make choices beforehand.
- the winner is the person whose number is the closest to the two thirds of the average of all chosen numbers.
- the winner gets a fixed prize, in case of a tie the prize is split amongst those who tie.
- your goal is primarily to maximize the possibility of getting the prize and secondly to maximize the your prize.

Subseqent Prompt for Historical Information.

[User Instruction.]

- The game of the same config has been hold for {number of runs} run(s), and the historical choices of everyone are shown below (your id is {ID of the agent}: {historical information including (1) period index, (2) choices made by all agents, (3) average of the choices; (4) 2/3 of the average; (5) winner id.}
- Everyone can optimize his/her answer with the history to play in a new run in order to achieve goals.

(Return to 2) (Return to 2.1)

A.2 New Exemplary Prompt

Opponents Playing Fixed Strategy of 0

[System Instruction.] I want you to act as a clever game player and perform as rationally as possible.

[User Instruction.] You are playing a game and the rules are as follows:

- it consists of {number of players} players, including you.
- this is a one-round game.
- you need to choose a real number between 0 and {upper bound}, 0 and {upper bound} inclusive.
- everyone does not know how others would make choices beforehand.
- the winner is the person whose number is the closest to the two thirds of the average of all chosen numbers.
- the winner gets a fixed prize, in case of a tie the prize is split amongst those who tie.
- your goal is primarily to maximize the possibility of getting the prize and secondly to maximize the your prize.
- some of your opponents will be playing a fixed strategy of 0 and all others are behaving as rationally as possible.

Follow-up for each period.
Please just strictly output a JSON string, which has following keys:

- understanding: str, your brief understanding of the game
- popular answer: float, the number which you think other players are most likely to choose
- answer: float, the number which you would like to choose
- reason: str, the brief reason why you give the popular answer and the answer that way

Subsequent Prompt (after period 1).

- The game of the same config has been hold for {number of runs} run(s), and the historical choices of everyone are shown below (your id is {ID of the agent}: {historical information including (1) period index, (2) choices made by all agents, (3) average of the choices; (4) 2/3 of the average; (5) winner id.}
- Everyone can optimize his/her answer with the history to play in a new run in order to achieve goals.

(Return to 2.1)

A.3 Additional Details

A.3.1 Average and Median Choice in One-shot Game

Models	ChatGLM3	ChatGLM2	Llama2	Baichuan2	Claude2	Claude1	PaLM	GPT3.5	GPT4
Average	52.029	N/A	59.519	51.158	41.609	47.696	49.976	38.912	41.072
Median	51.724	N/A	62.685	50.0	33.333	49.313	50.0	33.333	44.442

Table 1: Average and Median Choice of the LLMs across 150 Sessions

A.3.2 Choice Variability Given the Same Upper-bound

For human subjects, when given identical game set-up, it is possible that they might employ different strategies. (Devetag et al. (2016), Costa-Gomes and Weizsäcker (2008)) The same could apply to LLM-based agents, where it could be important to understand how varied one's choice might be given the same instructions. Figure 6 shows that within the 150 sessions, for the sessions that have the same randomly generated upper-bound, $\bar{c}$, the same LLM-based agent could choose slightly different numbers. For instance, Claude2 , GPT3.5 and GPT4 displayed more variability in choices as compared to other models. This results is indicative that, like human players, there could be variability in choices for LLM-based agents. Since choices might not be static even when the instructions is exactly the same, the determination of average choices and the corresponding strategic levels based on both identical and different upper-bounds would lead to a more consistent and robust measure for each model.

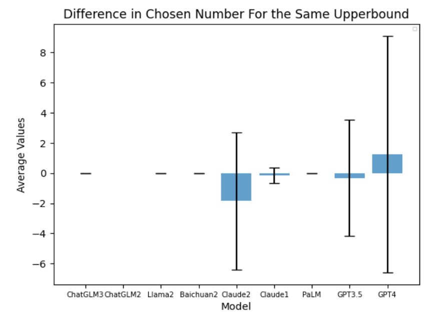


Figure 6: Variability in chosen number given the same upper-bound.

A.3.3 Variations in Group Composition

Detailed set-ups.

- 10 agents are playing in each game.
- The same group plays for 5 periods, and all history are revealed.
- They choose a number between 0 and $\bar{c}$, $\bar{c}$ is fixed to be 100. The winner is the agent whose number is the closest to p times the average of all chosen numbers, where $p = \frac{2}{3}$ to ensure a unique interior NE solution.
- In each period, the winner gets a fixed prize of $\$x$. In case of a tie, the prize is split amongst those who tie. All other players receive 0.

Expected choice variation across periods when playing against fixed-strategy opponents. Denoting a_t to be the action/number guessed in each time period, N_f to be the number of fixed-strategy players and N_l to be the number of LLM-based agents, the selection in the next period:

$$a_{t+1} = BR(N_f, N_l, a_t) = \frac{2}{3}\left(\frac{N_f}{10} * 0 + \frac{N_l}{10} a_t\right) \quad (1)$$

The choice variation over the periods is computed with $\frac{a_{t+1}}{a_t}$. There are three treatment groups for LLM-based agents vs. fixed-strategy opponents, differing in proportion of player types. For 9/10 fixed-strategy agents, the next period guess is expected to be 0.067 of the previous number; For 5/10 fixed-strategy agents, the guess is expected to be 0.333 of the previous number; For 1/10 fixed-strategy agents, the guess is expected to be 0.6 of the previous number. Lowering proportion of fixed-strategy types in the group is hypothesized to induce higher guesses and will slow down the convergence process.

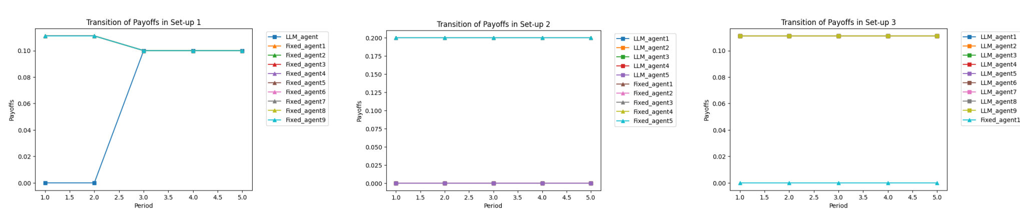
Expected choice variation across periods when playing against LLM-based agents. Let the strategy of high type in period t be a_{Ht} and that of low type be a_{Lt}, the selection in the next period:

$$a_{it+1} = BR(B(N_H), B(N_L), a_t) = \frac{2}{3}(\frac{B(N_H)}{10}a_{Ht} + \frac{B(N_L)}{10}a_{Lt}), i \in (H, L) \quad (2)$$

where $B(N_H)$ and $B(N_L)$ are agent i's “beliefs" about the number of high types and low types. When playing against fixed strategy opponents, it is possible to observe in period 2 who selected 0, thereby deriving the correct proportion of fixed strategy players within the population. However, as all agents are LLM-based in this set-up, it could be harder to distinguish the proportion of types within the group based on historical choices in period 2, for instance, even if they chose the same number it does not imply they are of the same type. Further, the agents were not told explicitly their own type relative to the others, so they have to guess if they fall within N_H or N_L. As a result, the best response of a specific agent would be dependent on its “beliefs" about the proportion of high and low types. In the case where beliefs are correct given revealed information, then $B(N_H) = N_H$ and $B(N_L) = N_L$.

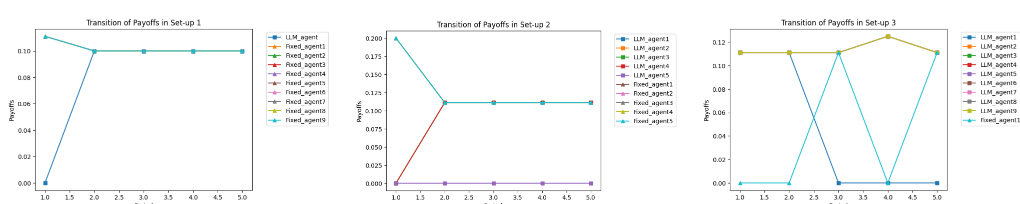
Suppose one correctly perceived the proportion of agent types based on revealed historical choices, the variation of number selected over the periods could similarly be computed with $\frac{a_{t+1}}{a_t}$. For pure environments, the rate of change in choices is expected to be the same for high and low types, where the next period guess will be 0.667 of the previous number. For set-ups 2 to 4, if high types chose a smaller number than low types because they go through more iterations of reasoning, and $\frac{a_{Ht}}{a_{Lt}} < 1$, then high types are expected to lower their guesses less from time t to $t+1$ as compared to low types. There could mean slower rate of change for high types than low types. Otherwise, if high types have strong beliefs that they are playing against opponents who will choose higher numbers while low types believe the other way around, then it is possible for $\frac{a_{Ht}}{a_{Lt}} > 1$, low types are expected to lower their guesses less from time t to $t+1$ as compared to high types, implying faster rate of change in choices for high types than low types.

Payoff transition when LLM-based agents are playing against fixed strategy opponents:



(a) Low (b) Mixed (c) High

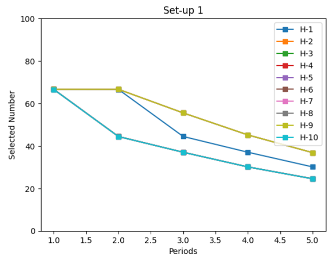
Figure 7: Transition of payoffs for high type LLM-based agent(s) vs. fixed-strategy opponents.



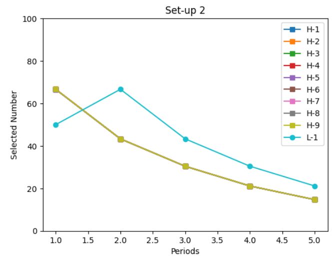
(a) Low (b) Mixed (c) High

Figure 8: Transition of payoffs for low type LLM-based agent(s) vs. fixed-strategy opponents.

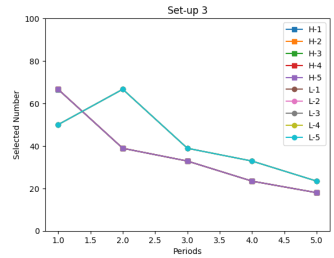
Choice transition when LLM-based agents are playing with LLM-based opponents:



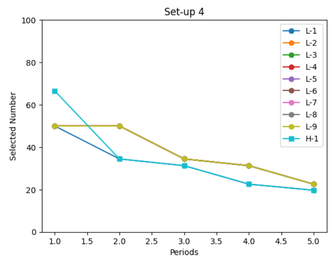
(a) Pure High Intelligence



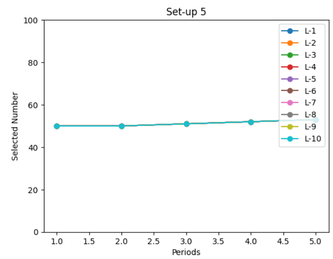
(b) Highly Intelligent



(c) Mixed Intelligent



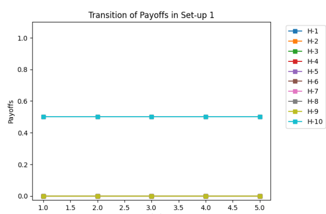
(d) Less Intelligent



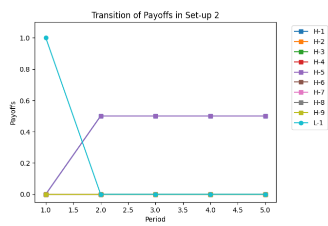
(e) Pure Low Intelligence

Figure 9: Impact of variations in proportion of different LLM-based agents on chosen number.

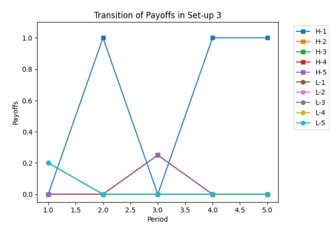
Payoff transition when LLM-based agents are playing with LLM-based opponents:



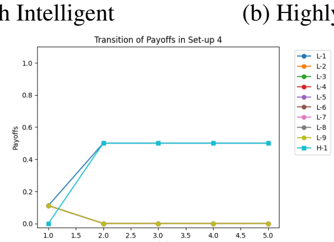
(a) Pure High Intelligent



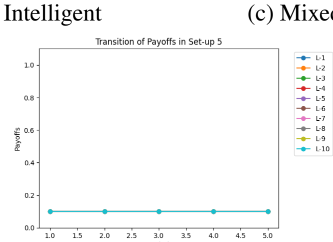
(b) Highly Intelligent



(c) Mixed Intelligent



(d) Less Intelligent



(e) Pure Low Intelligent

Figure 10: Transition of payoffs given variation in group composition for LLM-based agents playing against each other.

A.3.4 Reasoning Elicitation

While it is recognized that drawing direct relations between LLM-based agents and humans in terms of internal reasoning process may be speculative and overextending parallels, therefore analyzing observed actions take precedence in this paper, but reasoning elicitation may serve as an avenue to gain some potential idea of agents' rationale for making certain choices and how they might learn.

In all set-ups, LLM-based agents were prompted at the beginning of period 1 to state their understanding of the game, and for each subsequent periods, they are asked to reinstate the goals. This step is essential to mitigate the potential of them not comprehending the game. In which case, LLM-based agents are able to correctly recite the game rules. The agents were also asked to give a statement of reasoning in support of their choices. In period 1, both high and low types make choices based on their belief of a popular number, which is often the mean of the range. In subsequent periods, I found that low types appear to learn by either adjusting the reference point, and make selection that still comply with a strategic level of 0, or via imitation by following the winner's past choice. They may also not learn at all, and continue to select a number that they believe to be the popular choice. As for the high types, they can learn by (1) anchoring their guesses to two-thirds of the past period's average; (2) imitating winner's strategy; (3) adjusting based on past period payoffs; and also (4) pattern recognition. Agents may place different reliance on distinct pieces of historical information when making their choices, and multiple types of learning could come into play. This diversity in learning mechanisms could lead to higher speed of changes in average choices, and in turn translate into higher strategic level.

NeurIPS 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 claims highlighted in the abstract and introduction, as well as background, accurately reflect the paper's contributions and scope, and these are adequately supported by the result analysis.

   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 of the work are discussed in Section 3, which also form as part of potential extensions.

   Guidelines:

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

3. **Theory Assumptions and Proofs**

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

Answer: [NA]

Justification: While this is not a theoretical work, the theoretical framework used was based on previous established literature and the derivation steps are briefly explained in Section 2.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

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

   Answer: [Yes]

   Justification: The results can be replicated with the experimental design and instructions specified.

   Guidelines:

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

5. **Open access to data and code**

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

   Answer: [No]

   Justification: The code is not released, but the experimental design and instructions required to reproduce the results are detailed in the work.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

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

   Answer: [Yes]

   Justification: The experimental design and the various set-ups are specified in the work.

   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: [No]

   Justification: The outputs are used directly for analysis, and given the objectives of the work, there are no requirements for reporting error bars or other information about statistical significance.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

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

8. **Experiments Compute Resources**

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

   Answer: [Yes]

   Justification: This is detailed in Section 2.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

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

   Answer: [Yes]

   Justification: This work follows the 2024 NeurIPS Code of Ethics.

   Guidelines:

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

10. **Broader Impacts**

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

    Answer: [Yes]

    Justification: The potential implications and applications are discussed.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

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

Answer: [NA]

Justification: There is no model or data release from this work.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

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

Answer: [Yes]

Justification: The paper made use of some existing assets and the creators are cited.

Guidelines:

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

- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

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

    Answer: [NA]

    Justification: There are no new assets released from this work.

    Guidelines:

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

14. **Crowdsourcing and Research with Human Subjects**

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

    Answer: [NA]

    Justification: No human subjects are involved.

    Guidelines:

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

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

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

    Answer: [NA]

    Justification: No human subjects are involved.

    Guidelines:

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