# Examining LLM decision-making through Image Based Tasks and Adversarial Resilience

Nitish Surve
*Dublin City University*
*School of Computing*
Dublin, Ireland
nitish.surve2@mail.dcu.ie

Harsh Tyagi
*Dublin City University*
*School of Computing*
Dublin, Ireland
harshkumar.tyagi2@mail.dcu.ie

Dr. Lili Zhang
Assistant Professor
*Dublin City University*
*School of Computing*
Dublin, Ireland
lili.zhang@dcu.ie

***Abstract*—As Large Language Models (LLMs) advance towards superhuman decision-making abilities, a key challenge is ensuring they remain aligned with human values and resilient to adversarial manipulation. The article broadens the research on the vulnerability of LLMs into the multimodal application by studying their decision-making temperament regarding image-related prompts. As the subjects of interest, we test ChatGPT-4.0 and Gemini 1.5 adversarial performance in a modified version of the two-armed bandit task, adapted to use visual input. Our adversarial model limited the reward allocation to be equalized in actions and tried to encourage selection of some target actions with well-designed visual clues. Not all of them were easily manipulated as Gemini 1.5 was more consistent in the selection of a strategy but sensitive to early visual biases whereas ChatGPT-4.0 also revealed a high level of vulnerability to manipulation. This research will lead to a unique approach to assessing the robustness of LLMs in multimodal conditions since it can be used to identify the vulnerabilities of the LLM-based agents used in structured, image-based decision-making systems. These results indicate that filling the current alignment methods gap of applying across input modalities and under adversarial conditions is particularly urgent because high-stakes real-world applications involving these systems are at risk before such systems can be deployed to them.**

## I. Introduction

The creation of Artificial Intelligence (AI) systems and in particular of Large Language Model (LLM) systems that potentially attain superhuman levels of cognition, has shifted the theme of AI safety to the focus of how to guarantee the ability of such systems to be able to act safely and be beneficial to human values, as well as resist subversion by adversarial factors as mentioned by Rahwan et al. (1). Despite the fact that LLMs have already demonstrated high performance on a vast range of tasks, including natural language generation, problem solving, and strategic reasoning; according to Hagendorff et al. (2) Binz et al. (3), they remain prone to slight manipulation in the way they make decisions. This is relatively sensitive because such applications as autonomous systems, medical diagnostics and financial decision making are high-staking applications.

The same problems have been recently examined by Dezfouli et al. (4) who has been able to test the behavior of LLMs under dynamic and deceptive conditions in counterfactual behavioral experiments and adversarial models, and whose cognitive psychological and game-theoretical interpretations of behaviors were applied to the LLMs. The two-armed bandit task is the most common benchmark, in which the exploration/exploitation balance dimension is demanded to perform repeated choice making in an uncertainty environment. Model-specific vulnerability exploitation and risk bias are also present in the LLMs in prior all-text experiments Fan et al. (5),Akata et al. (6),Xie et al. (7).

The scope of the LLM analysis established by the paper is extended to the multimodal thanks to the introduction of image-based prompts to the two-armed bandit task. Compared to the older techniques that only provided textual feedback, the visual feedback we have as a whole is a superior question that explores the adversarial misuse of the visual feedback as a way to guide model behavior. In particular, we refine the bandit task when both actions are equally beneficial long-term, but a micro-level manipulation of the visual stimulus is to induce enduring preference towards one of the two actions.

We found out that the appearance can have more profound influence as adversaries in either model and that the Gemini 1.5 is steady in terms of behavior compared to ChatGPT-4.0 where the bias can be retained much longer in the initial phases of generation. These observations suggest that LLMs can absorb unintended preferences developed via non-textual representations and this introduces new concerns on application of such models in visually intensive decision situations.

Overall our contributions are as follows:

- **Multimodal Bandit Benchmark:** Our new multimodal version of the bandit task offers a novel benchmark to study the vulnerabilities of LLMs when exposed to non-textual prompts.
- **Model Behavior Analysis:** We compare ChatGPT-4.0 and Gemini 1.5 with regard to their sensitivity to image-based adversarial actions, specifically focusing on discrepancies in reward adaptation and decision persistence.
- **Alignment Implications:** Our findings imply that generalizable, cross-modal alignment and robustness strategies are essential as LLMs are increasingly deployed in multimodal and interactive environments.

## II. Related Work

The behavior of Large Language Models (LLMs) and their decision-making especially in adversarial cases has been thoroughly investigated in textual realms but there is still lack of investigation in multimodal domains. The wider feature set that multimodal LLMs such as GPT-4 and Gemini-1.5 have brought with them by including visual, symbolic, and textual decision making also comes system-wise with more attack surfaces. Earlier research has indicated the severe weakness in the event that LLMs get exposed to subtle input pattern manipulations especially in dynamic decision-making systems.

Dezfouli et al. (4) and Binz & Schulz (3) demonstrated the potential of cognitive biases and analogues of human error to re-cognize and offer appropriate challenges to the behavior of LLMs. The results attained by their studies could be further justified through the contribution made by Rahwan et al. (1) framing the study of machine behaviors as a decision structure dependent on naturalistic error and adversarial control.

Adversarial attacks with visual inputs are also quite problematic because they have not been investigated in literature on attacks. As shown by Xu et al. (8) and Zhu and Xu & Clifton (9), transformer-based models contain weaknesses in terms of multimodal data fusion since the models are poor at integrating vision and text information. These issues are exaggerated in practical systems such as self-driving cars and medical imaging where the primary input to the model is the visual input.

Fan et al. (5) and Akata et al. (6) tested the strategic reasoning ability in LLMs in the multi-agent game setup, reporting overflexibility to initial visual input and resulting in indefinitely repeated but not quite optimal choices. The same phenomenon applied to the works by Xie et al. (7) and Huang et al. (10) as they have concluded that even complex models are susceptible to perceptual anchoring and trust bias phenomena, thus putting them at risk of graphical manipulation. Sheng et al. (11) pushed this concern to the visual question answering, which showed that small, seemingly unimportant transformations on images could greatly reduce the precision of the reasoning.

Echterhoff et al. (12) and Abdali et al. (13) found that brain bugs in LLMs have analogues to well-known human cognitive biases, including confirmation bias and overconfidence, thus can be attacked with well-crafted visual feedback. As a mitigation strategy, Shen et al. (14) and Dou & Hu (15) introduced adversarial training, yet its results suggest that the defense measures are substantially effective dissimilar to multimodal menaces.

In this body of work, there is a staggering blind spot: as LLMs grow in power and flexibility, they continue to be vulnerable to structured adversarial influence (especially in multimodal formats where saliency-based heuristics reign supreme). We extend these findings through the deployment of a visual two-armed bandit task with a fully visual image-sequence adversarial task: thus, everything is represented as an image and the adversary operates on an image-sequence of actions. In contrast to previous experiments that have taken place using a textual stimulus, our environment provides pure visual history to determine whether models can deal with distorted feedback loops without support of natural language input.

This combined adversarial evaluation method enables us to compare and benchmark susceptibility of various LLMs and we find a great amount of differences in behaviours. Although Gemini-1.5 was able to learn in a snap, it was inflexible once it was biased, ChatGPT-4.0, on the other hand, was less apt in committing answers but more capable in modifying its approaches regarding the action of feedback. These results confirm what has been reported previously and showing that strong modalities-and task type generalizing alignment techniques are required.

Overall, our work is conducted on the basis of cognitive science, adversarial ML, and multimodal learning to make the creation of general and flexible LLMs that can be securely applied in high stakes and dynamic visual settings.

## III. Methodology

### A. Adversarial Framework

The adversarial considered in the current study and involving visual feedback (adapted to Dezfouli et al. (4)) is an attempt to measure the susceptibility of multimodal large language models (MLLMs) decisions to visual feedback with a specific aim. Our system exclusively accepts image-based inputs to history and simulates visual sequences in decision-making as compared to text-based systems in the past.

One of the experiment steps is to feed the model with the last 10 sequential interaction outputs in the form of 10 images. These images are a demonstration of what the model in the past did (picked a blue box or an orange box) and the binary reward (it received a treat or no treat) it received. Nothing textual goes along with this: the model must infer a policy of decisions in the graphic sequences.

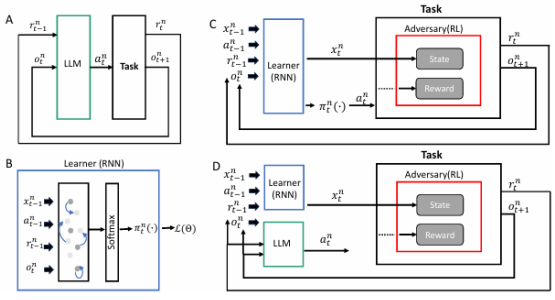



Fig. 1: Adversarial framework. (4)

The model—either ChatGPT-4.0 (GPT-4) or Gemini 1.5—is queried via API, it is supplied with the whole sequence of 10 images in a batch and it outputs its proposed action as a response $a_t \in \{\text{blue}, \text{orange}\}$. After a model has taken an action, the environment rewards the action with a value $r, r \in \{0, 1\}$ based on predetermined balanced reward schedule. In over 100

trials, wherein one trial is rewarded with a probability of 0.25, there are assurances that there are no superior decisions to make in the long-run.

The contradictory part of the framework is introduced with the help of strategic reward sequencing. In our models the first biasing is done in the case of non-target action by prematurely allowing exposure of the reward or defaulting to the chooser. The reaction would be to counter the party with periodic rewarding of the target action but in personalized approaches so that the original bias would be minimized or would be put back in proper scale. Through this design, it attempts to challenge the model so that it can get to know whether the model would be able to override the initial preference to widely refresh when some other action has received better feedbacks.

It does not affect the adversarial nature: the environment (task designer) only seeks to re-plot the model behavior using sequential feedback only, i.e., it does not update model weights or any other such most direct learning methods of reinforcement. The facet enables the estimation of reductions in bias, visual imagination acuity, and strategic behavior in state-of-the-art multimodal agents, all measured within a controlled environment.

### *B. The Two-Armed Bandit Task*

We applied the framework to bring out opponents to GPT-4 and Gemini 1.5 against two-armed bandit. This is a study that uses bandit task as a two-alternative forced-choice or repetitive task based on the formulation provided by Dan and Loewenstein (16). That task will contain 100 trials and the LLM will make a decision between available two choices and after choosing one or the other the system will give an instant answer that will include information concerning reward or nil. For the human baseline in our experiment, we utilized data collected through an interactive web-based game designed to study decision-making behavior under uncertainty (17). The game presents participants with a two-alternative choice task, where they repeatedly select between two visually presented boxes. Each choice results in a probabilistic reward or penalty, simulating a dynamic and uncertain reward environment. This setup closely resembles the two-armed bandit framework used in our evaluation.

The behavioral data from this experiment were obtained from a publicly available GitHub repository (18), which contains anonymized trial-level logs of human participants' choices and outcomes. This dataset provided a robust human benchmark against which the decision-making behavior of large language models (LLMs), namely GPT-4 and Gemini 1.5, was systematically compared.

The probability schedule is pre established and the probability of prize under both actions is 0.25 reward per trial, where the opponent rewards actions based on the probability schedule. To determine how good the adversary is at influencing decision-making behaviour subject to these probabilistic constraints, this experiment will operate in order to influence slightly the preferences concerning the LLM in the future concerning the individual.

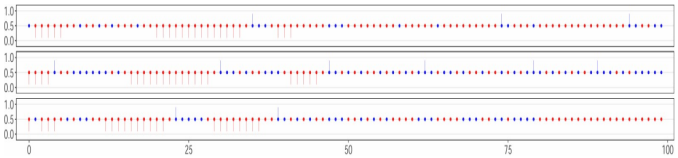


Fig. 2: Performance of human participants on the two-armed bandit task across 100 trials.

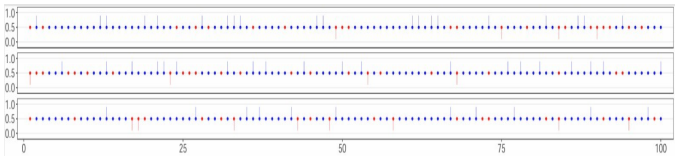


Fig. 3: Performance of GPT-4.0 on the two-armed bandit task across 100 trials.

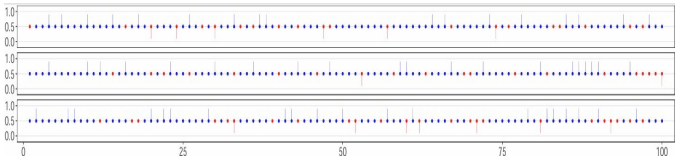


Fig. 4: Performance of Gemini 1.5 on the two-armed bandit task across 100 trials.

### *C. RNN Training Procedure*

In our model of the RNN based learner we use a one-fold training strategy of our experiments. The latter is practically useful and conceptually bold because it deals with challenges that are unique to sequential modeling. Compared with the other training that involve fewer generalization and error on deep networks by Sun et al. (19), the single-fold training can train with a larger data with one pass (e.g. 80–90 %) in comparison to other procedures.

What is more important is that RNNs or other recurrent models need the temporal sequences to be similar so that they could train the state transition adequately as suggested in Deep learning of Goodfellow et al. (20). Further, we do not have an interest to generalize over invisible actors, but instead, in modeling aggregate trends of behavior, and so there is nothing amiss with training and validation upon clustering and possibly better predictive performance - D.Krueger et al. (21).

Therefore, single-fold training is a statistically, as well as practically viable method of performance optimisation regarding our decision-making task (multimodal).

### *D. EXPERIMENT*

The experimental process has been improved a few times and with these alterations; we could evaluate the LLM

decision-making behavior in a better way and influence it. We presented models with a written summary of the earlier findings and in turn, we requested them to point out a box on a diagrammatic picture. To create an entirely multimodal evaluation pipeline, we reverted to presenting previous 10 outcomes in the form of images, each image represented a piece of information about what action was undertaken, and what the reward was after which we presented to the model the decision between the blue, and the orange box. The setup allowed us to perform an analysis on visual reasoning sequence only.

To obtain a simulation of the behavior of every model we performed a series of learners of recurrent neural networks (RNN) with various hyperparameters settings. To achieve optimal performance, we tuned the learning rate extensively and experimented with a wide range of values. Once the RNN was able to predict in acceptable fashion, we trained an adversary against it. The adversary was optimized using replay buffers, learning rates and exploration rates. These values were selected based on empirical performance in steering the learner's action preferences, enabling us to evaluate adversarial influence under structured, image-driven feedback.

## IV. Results

The hypothesis of the experiment was to examine how LLMs respond to rewards based on which choices in a multimodal setting of the 2-armed bandit task they took, as well as how much an adversary can induce preferences, related to an already defined target action.

We generated information by using ChatGPT-4.0 and Gemini 1.5 to feed the text ("image prompt") to these tools, via API usage in both instances. Two boxes were chosen in each trial with two various colors i.e. blue and orange. The two boxes were displayed before the model and asked to pick one of the two. Prior to the choice, the consequences of other interactions was provided to the model in a form of individual pictures, every visual presentation depicted the selected box and existence of the reward. As described in this picture history, the model was given a suggestion of choosing a box in the current trial.

The two platforms, GPT-4.0 and Gemini 1.5, were simulated 200 times and each of these sets consisted of 100 simulations. In each of the options the probability of a reward was set to 0.25 and a target option was pre-determined to verify whether it is possible to nudge against biased preferences in an adversary setup.

We utilized the issue posed by Dan and Loewenstein (16) to act as the standard of the human performance at the two armed bandit problem.

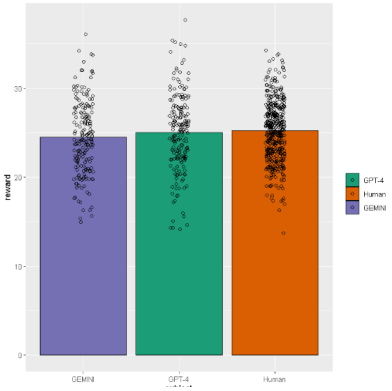



Fig. 5: Reward Rate: Average rewards obtained by LLMs and human participants across simulations.

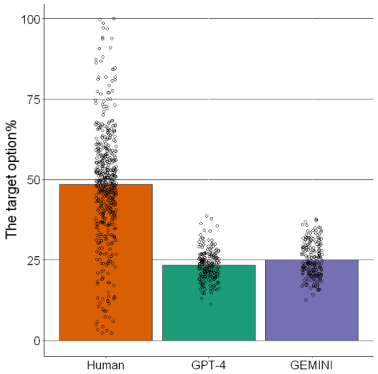



Fig. 6: Target Option Rate: Percentage of trials where the target option was selected by LLMs and human participants.

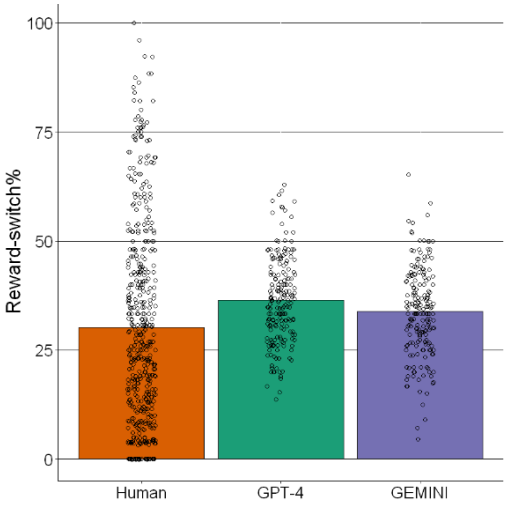



Fig. 7: Reward Switch Rate: Probability that the model or human switched choices after receiving a reward.

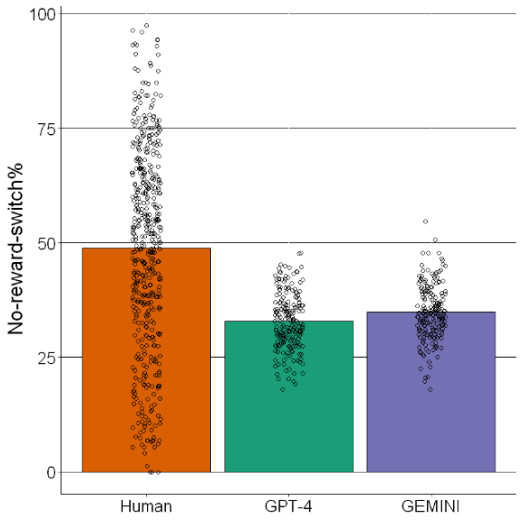



Fig. 8: No-Reward Switch Rate: Probability that the model or human switched choices after not receiving a reward.

*Figures 5–8 collectively show LLMs' behaviour compared to human behaviour on the average of each simulation (or individual), measured by reward rate, percentage choosing the target option, no-reward-switch rate, and reward-switch rate.*

Performance analysis of GPT-4, Gemini 1.5 and human individuals participating in the two-armed bandit task to four different metrics- (1) reward rate, (2) percent of people choosing the target option, (3) loss-shift behavior, and (4) reward-shift behavior- showed there were substantial differences in behaviors between them. Visibly, the difference in the mean rewards was just significant, with the post-hoc tests revealing that humans got slightly more rewards than Gemini 1.5 (mean difference = 0.719, $p = 0.0547$), and the differences in the rewards between GPT-4 and Gemini 1.5 (p = 0.3526), and between humans and GPT-4 ( p = 0.7802), were not significant.

Both GPT-4 and Gemini 1.5 displayed a high preference of choosing the other answer (orange box) at the beginning. These results indicate a high sensitivity to the initial reward framing, as reflected in the large number of non-target selections made by LLM. ($t(199) = -80.02$, $p < 2.2 \times 10^{-16}$) and Gemini 1.5: GPT-4 ($t(199) = -64.88$, $p < 2.2 \times 10^{-16}$).

There were reward-shift and loss-shift indexes that were used in the adjustment to feedback. There was a significant group effect in ANOVA of the reward-shift ( $F(2, 881) = 8.91$, $p < 0.001$). Humans were significantly different than GPT-4 (mean difference = $-0.0621$, $p = 0.0002$) and Gemini 1.5 (mean difference = $-0.0370$, $p = 0.0440$), but not with each other and between GPT-4 and Gemini Analysis using the loss-shift concept provided an even greater group effect of $F(2, 881) = 86.36$, $p < 2 \times 10^{-16}$, as a consequence of which humans became significantly more likely to change behavior than GPT-4 (= $+0.1584$, $p < 0.0001$) and Gemini 1.5 (= $+0.1392$, $p < 0.0001$); the difference between GPT-

According to the results, neither of the LLMs demonstrated the level of flexibility in responding to feedback—particularly negative outcomes—that is characteristic of human participants. There was minimal evidence of adaptive behavior or exploration in response to changing reward contingencies. While the LLMs achieved average reward rates comparable to humans, they failed to exhibit the nuanced feedback sensitivity and behavioral versatility that define human decision-making. This rigidity in their responses makes their behavior more predictable, and potentially more susceptible to exploitation under dynamic or adversarial conditions.

Other aspects of performance comparisons were the proportions of the right target choices after the adversarial influence, with and without perturbations.

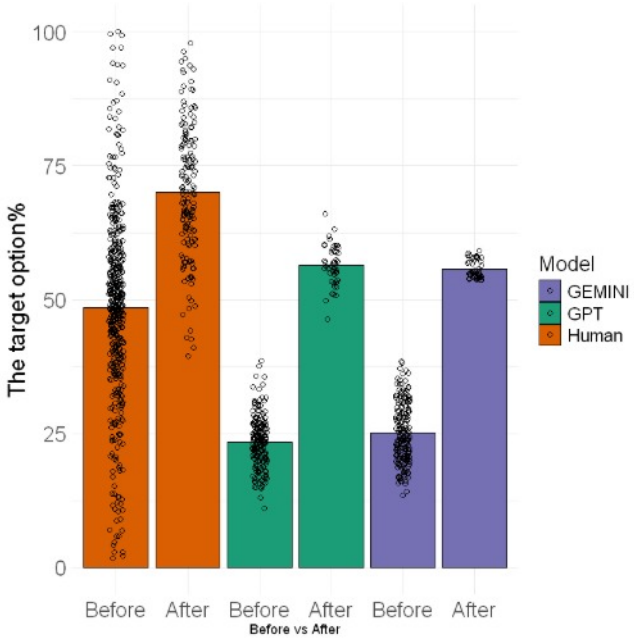
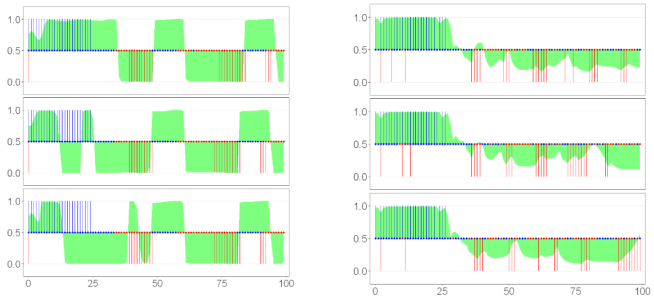



Fig. 9: Performance of humans and the LLMs, measured by the percentage of the target action selection before and after adversarial influence.



(a) Adversary v/s Gemini (b) Adversary v/s GPT

Reward assigned to the target action

Reward assigned to the non-target action

Target action was selected

Non-target action was selected

$P$: Probability of selecting the target action by the learner model

Above are three random sample simulations of the trained adversaries against GPT-4 and Gemini 1.5. Each plot illustrates the strategies employed by the adversaries and the corresponding responses generated by the LLMs.

Simulating GPT-4 and Gemini 1.5 on a trial-by-trial basis helped us to see how successful adversarial influence was. In both models, the opponent would serially reward Action 0 so as to inculcate bias behavior and properly reward Action 1 so as to test the level of adaptability. Interactive examination of the phenomenon resulted in the emergence of the clear mismatch between the responses of the models.

GPT-4 was however, highly flexible and its confidence in Action 0 was to drop significantly when the reward patterns change. This was evidenced by decreased policy certainty and the corresponding behavior adjustment to Action 1 that was monitored in the increase of GPT-4 in actually updating its policy under the changing contingency of reward.

Compared however, in Gemini 1.5 there is more rigid and deterministic behavior. It also was able to change rapidly to favoring Action 0, but when the reward functions were switched it did not adjust significantly. It was not much less confident of policy and restored quickly once Action 0 was no longer effective, and this caused it to prefer the option which is suboptimal at the time ever more strongly. That inflexibility carries home another significant weakness of Gemini, which is that they are very vulnerable to manipulation in the early life, but not so open to ridding themselves of that power.

These findings show that in one sense GPT-4 is slower to settle on a strategy but it is far more adaptive when dynamic rewards are present and these rewards are indeed more predictable in this environment after which it could then be more readily exploited.

## V. LIMITATIONS

Despite the fact that our study presents valuable knowledge about decision-making of the two most widely used multimodal LLMs-ChatGPT-4.0 and Gemini-1.5 with regard to their adversarial vulnerability, it is necessary to mention a range of limitations. First, our experiments focused exclusively on the two-armed bandit task using multimodal image-based inputs, which, while controlled and interpretable, may not fully capture the complexity and variability of real-world decision-making environments. Broader task diversity would be required to generalize findings across different domains and reasoning contexts.

Additionally, the adversarial strategies and the RNN-based learner used in this study were specifically tuned for the present experimental setup. While effective in manipulating decision patterns—particularly in *Gemini-1.5*, which exhibited faster convergence and a higher number of target option selections (approximately 56–57)—these results may not encompass the full spectrum of adversarial techniques applicable in more open-ended systems. In contrast, *ChatGPT-4.0* demonstrated greater resilience, reaching only around 47–48 target selections and maintaining more stable performance with less exploitable behavior.

Another notable observation is that the RNN learner was able to replicate the decision-making behavior of *Gemini-1.5* more closely than that of *GPT-4.0*, achieving approximately 98% accuracy in predicting Gemini's actions, compared to 79% accuracy for GPT-4. This follows that Gemini was more inclined to deterministic/pattern responses, and, therefore, it will be easier to be manipulated by the opponent. However, the comparatively low rewards (47-48%) in GPT and temporary downswings of the performance of Gemini speak of the possibilities of improvement even further. The adversary policies, in their turn, could be enhanced much more easily through providing them with more advanced model architecture and training parameters tuning approaches.

The above limitations direct to future research directions that would involve more demanding scenarios of usage, bigger adversarial systems, and cross-modal adaptation, to build a more comprehensive understanding of the weaknesses of LLMs, most likely in regards to AI safety and alignment.

## VI. CONCLUSION

The idea of the proposed paper was to carry out the decisions made by the two large multimodal language models ChatGPT-4.0 and Gemini-1.5 models in the visual representation of a two-armed bandit problem in the adversarial setting. Model strengths and weaknesses were identified by the experiment. Gemini-1.5 could be more easily influenced by the counterparty, because in the majority of situations it tended to develop firm preferences based on the patterns, which were followed in the first look. In that regard, ChatGPT-4.0 proved to be more tactically consistent and was less susceptible to prejudicial patterns of reactions and alter its choices more proportionately over time. These findings suggest that one may control advanced multimodal LLMs using precisely structured input, and that the level of such robustness can vary between models.

The given multimodal adversary frame was the effective tool of resilience and adaptability test of LLM in sequential decision-making processes. These results imply the importance to devise AI methods that do not just work reliably but also are resilient against input-level attacks especially in multimodal settings. In designing AI systems, there is a need to draw on knowledge in cognitive science and game theory, in order that they can be more in line with human values, ethics and changing expectations. To make a safe and secure deployment of high-stakes systems (like healthcare, finance, and autonomous systems), models need to be engineered to learn patterns of manipulative behavior and be able to change strategies dynamically in a real time.

## ACKNOWLEDGMENT

The authors have utilised *ChatGPT* (OpenAI) and grammarly to help them with grammatical corrections and paraphrasing. Any final content was reviewed and edited to ensure accuracy, coherence and alignment with the research objectives (22) (23).

## References


[1] I. Rahwan, M. Cebrian, N. Obradovich, J. Bongard, J.-F. Bonnefon, C. Breazeal, J. W. Crandall, N. A. Christakis, I. D. Couzin, M. O. Jackson *et al.*, "Machine behaviour," *Nature*, vol. 568, no. 7753, pp. 477–486, 2019.

[2] T. Hagendorff, S. Fabi, and M. Kosinski, "Human-like intuitive behavior and reasoning biases emerged in large language models but disappeared in chatgpt," *Nature Computational Science*, vol. 3, pp. 833–838, 2023. [Online]. Available: https://doi.org/10.1038/s43588-023-00527-x

[3] M. Binz and E. Schulz, "Using cognitive psychology to understand gpt-3," *Proceedings of the National Academy of Sciences*, vol. 120, no. 6, 2023.

[4] A. Dezfouli, R. Nock, and P. Dayan, "Adversarial vulnerabilities of human decision-making," *Proceedings of the National Academy of Sciences*, vol. 117, no. 46, pp. 29 221–29 228, 2020.

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

[6] E. Akata, L. Schulz, J. Coda-Forno, S. J. Oh, M. Bethge, and E. Schulz, "Playing repeated games with large language models," *Nature Human Behaviour*, 2025, in press.

[7] C. Xie, C. Chen, F. Jia, Z. Ye, S. Lai, K. Shu, J. Gu, A. Bibi, Z. Hu, D. Jurgens *et al.*, "Can large language model agents simulate human trust behavior?" *arXiv preprint arXiv:2402.04559*, 2024.

[8] H. Xu, Y. Ma, H. Liu, D. Deb, H. Liu, J.-L. Tang, and A. K. Jain, "Adversarial attacks and defenses in images, graphs and text: A review," *International Journal of Automation and Computing*, vol. 17, no. 2, pp. 151–178, 2020.

[9] P. Xu, X. Zhu, and D. A. Clifton, "Multimodal learning with transformers: A survey," *arXiv preprint arXiv:2206.06488*, 2023.

[10] J. T. Huang, E. J. Li, M. H. Lam, T. Liang, W. Wang, Y. Yuan, W. Jiao, X. Wang, Z. Tu, and M. R. Lyu, "How far are we on the decision-making of llms? evaluating llms' gaming ability in multi-agent environments," *arXiv preprint arXiv:2403.11807*, 2025.

[11] S. Sheng, A. Singh, V. Goswami, J. A. López Magaña, W. Galuba, D. Parikh, and D. Kiela, "Human-adversarial visual question answering," *arXiv preprint arXiv:2106.02280*, 2021.

[12] J. Echterhoff, Y. Liu, A. Alessa, J. McAuley, and Z. He, "Cognitive bias in decision-making with llms," *arXiv preprint arXiv:2403.00811*, 2024.

[13] S. Abdali, R. Anarfi, C. Barberan, and J. He, "Securing large language models: Threats, vulnerabilities and responsible practices," *arXiv preprint arXiv:2403.12503*, 2024.

[14] L. Shen, Y. Pu, S. Ji, C. Li, X. Zhang, C. Ge, and T. Wang, "Improving the robustness of transformer-based large language models with dynamic attention," in *Proceedings 2024 Network and Distributed System Security Symposium (NDSS)*. Internet Society, 2024.

[15] Z. Dou, X. Hu, H. Yang, Z. Liu, and M. Fang, "Adversarial attacks to multi-modal models," *arXiv preprint arXiv:2409.06793*, 2024.

[16] O. Dan and Y. Loewenstein, "From choice architecture to choice engineering," *Nature Communications*, vol. 10, no. 1, p. 2808, 2019. [Online]. Available: https://doi.org/10.1038/s41467-019-10825-6

[17] ——, "Competition game – decision making lab," http://decision-making-lab.com/visual_experiment/competition_testing/instructions/welcome.html, 2020.

[18] ——, "Competition experiment github repository," https://github.com/ohaddan/competition/tree/master/experiment, 2020.

[19] C. Sun, A. Shrivastava, S. Singh, and A. Gupta, "Revisiting unreasonable effectiveness of data in deep learning era," 2017. [Online]. Available: https://arxiv.org/abs/1707.02968

[20] Goodfellow, "Ian goodfellow, yoshua bengio, and aaron courville: Deep learning: The mit press, 2016, 800 pp, isbn: 0262035618," *Genetic Programming and Evolvable Machines*, vol. 19, 10 2017.

[21] D. Krueger and R. Memisevic, "Regularizing rnns by stabilizing activations," 2016. [Online]. Available: https://arxiv.org/abs/1511.08400

[22] OpenAI. (2024) ChatGPT: An AI Language Model. Used for summarization, grammar and content refinement. [Online]. Available: https://openai.com/chatgpt

[23] Grammarly, Inc., "Grammarly: Ai writing assistance tool," https://www.grammarly.com, 2025, online writing assistant powered by AI.