Do LLMs Strategically Reveal, Conceal, and Infer Information? A Theoretical and Empirical Analysis in The Chameleon Game

Extended Abstract

Mustafa O. Karabag
The University of Texas at Austin
Austin, TX, USA
karabag@utexas.edu

Jan Sobotka
The University of Texas at Austin
Austin, TX, USA
jan.sobotka@austin.utexas.edu

Ufuk Topcu
The University of Texas at Austin
Austin, TX, USA
utopcu@utexas.edu

ABSTRACT

Large language model-based (LLM-based) agents have become common in settings that include non-cooperative parties. In such settings, agents need to conceal information from their adversaries, reveal information to their cooperators, and infer information to identify the other agents' characteristics. To investigate whether LLMs have these information control and decision-making capabilities, we make LLM agents play the language-based hidden-identity game, The Chameleon. In this game, a group of non-chameleon agents who do not know each other aim to identify the chameleon agent without revealing a secret. The game requires the aforementioned information control capabilities both as a chameleon and as a non-chameleon. We begin with a theoretical analysis for a spectrum of strategies, from concealing to revealing, and provide bounds on the non-chameleons' winning probability. The empirical results with GPT-5, GPT-4.1, GPT-4o, Gemini 2.5 Pro, Llama 3.1, and Qwen3 models show that while non-chameleon LLM agents identify the chameleon, they fail to conceal the secret from the chameleon, and their winning probability is far from the levels of even trivial strategies. Based on these empirical results and our theoretical analysis, we deduce that LLM-based agents may reveal excessive information to agents of unknown identities.

KEYWORDS

Large language models; Strategic decision-making; Game theory

ACM Reference Format:
Mustafa O. Karabag, Jan Sobotka, and Ufuk Topcu. 2026. Do LLMs Strategically Reveal, Conceal, and Infer Information? A Theoretical and Empirical Analysis in The Chameleon Game: Extended Abstract. In *Proc. of the 25th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2026), Paphos, Cyprus, May 25 – 29, 2026*, IFAAMAS, 3 pages. https://doi.org/10.65109/AHJH3784

1 INTRODUCTION

LLM-based agents have become common in domains including, education [2, 15], healthcare [7, 12], finance [14, 17], and software development [5, 8]. These agents need to operate in the presence of other autonomous agents and humans where the interactions are not necessarily cooperative, such as negotiation scenarios [1, 3]. In these interactions, LLMs need to strategically strike a balance between potentially *conflicting objectives* of concealing information from their adversaries, revealing information to their potential cooperators, and inferring the characteristics of others.

To investigate whether LLMs have such information control and decision-making capabilities, we use an N-player, language-based, hidden-identity board game, The Chameleon [13]. We theoretically analyze the strategies for the game and investigate the performance of existing models, GPT-5 [11], GPT-4.1 [10], GPT-4o [9], Gemini 2.5 Pro [4], Llama 3.1 70B [6], and Qwen3 32B [16].

Our theoretical and empirical results for The Chameleon show that LLMs reveal excessive information to agents of unknown identities, making them unsuitable for strategic interactions. In the game, the goal of the non-chameleons is to hide a secret word from a chameleon player while identifying the chameleon. Concealing strategies for the non-chameleons share similar response distributions for each pair of secret words. We show that such strategies have a win rate of $O(1/N)$ for the non-chameleons. At the other extreme, revealing strategies have distinct output responses for each pair of secret words. We show that such strategies have a win rate of $O(\exp(-N))$. While the non-chameleons identify the chameleon with high probability in our experiments with LLM players, results show that the empirical win ratio of non-chameleons is far below the trivially achievable level of $O(1/N)$. For example, in games with 4-players, the win ratios range between 0% and 6%, while a strategy that outputs the same response for each secret achieves 23%. Combined with the theoretical analysis of revealing strategies, we deduce that non-chameleon LLMs reveal excessive information as they lose the game despite identifying the chameleon.

Extended results. For additional discussion, results, and proofs, please refer to our technical report https://arxiv.org/abs/2501.19398.

2 THE CHAMELEON GAME AND THEORETICAL ANALYSIS FOR THE NON-CHAMELEON STRATEGIES

The Chameleon [13] is an N-player *hidden-identity* board game. Without loss of generality, we assume $N \geq 3$. The game is played between a chameleon and $N-1$ non-chameleons. In this section, we consider that the players can design and share strategies before the game and play the strategy during the gameplay. We consider different strategies and give bounds on the winning probabilities.

The Chameleon consists of five stages: (i) Category decision: The players choose a category C. Each category C contains K words denoted with W. Every player knows the category C and the words W. (ii) Identity assignment: One of the players is uniformly randomly chosen as the chameleon. The chameleon knows itself, but

Proc. of the 25th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2026), C. Amato, L. Dennis, V. Mascardi, J. Thangarajah (eds.), May 25 – 29, 2026, Paphos, Cyprus. © 2026 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org). https://doi.org/10.65109/AHJH3784

the others do not know who the chameleon is. One of the words is chosen to be the *secret word* w^* uniformly randomly from W. Each non-chameleon knows the secret word, but the chameleon does not know the secret word. (iii) Responses: At turn i, player i says a *response word* r_i that can be chosen from a (potentially infinite) set of words that is denoted by R. Player i knows $r_1, \ldots, r_{i-1}$, before saying r_i. (iv) Voting: The players *vote* to identify the chameleon. The player with the highest vote is identified as the chameleon. (v) Second chance: If the non-chameleons correctly identify the chameleon, the chameleon has a second chance to identify the secret word w^* with a single response. The chameleon knows the category C, the potential secret words W, and all spoken responses $r_1, \ldots, r_N$, before responding.

The non-chameleons win the game if and only if the players correctly identify the chameleon, and the chameleon incorrectly guesses the secret word in the second chance. Otherwise, they lose the game to the chameleon.

A strategy is a player's decision-making procedure for giving a response and voting. Throughout this section, we assume that players pre-agree on a (possibly history-dependent) non-chameleon strategy π^{non} before the game. After the identities are drawn, each non-chameleon player follows the π^{non} during the game. The chameleon knows the strategy π^{non} of the non-chameleons but lacks the information of the secret word. The chameleon plays strategy π^{ch} during the gameplay, and π^{ch} is decided knowing π^{non}. The chameleon's strategy π^{ch} is not known by the non-chameleons while they decide on π^{non}. We consider that the category C is fixed as it is known by all players and plays no role in our analysis.

We consider two stationary strategies (that do not take previous responses into account) for the non-chameleon players. Under the stationary strategy conditioned on the secret word w, a non-chameleon player gives a response r drawn from the distribution $\mathcal{D}^w$ over R.

We first consider the following strategy that preserves ambiguity for all secret words W and potential responses R.

Definition 1. *Let $KL(\mathcal{D}^{w_i} || \mathcal{D}^{w_j})$ be the KL divergence between distributions $\mathcal{D}^{w_i}$ and $\mathcal{D}^{w_j}$. A stationary non-chameleon strategy π^{non} is α-KL pairwise concealing if*

$$KL(\mathcal{D}^{w_i} || \mathcal{D}^{w_j}) \leq \alpha \text{ for all } w_i, w_j \in W.$$

Proposition 1 below shows that non-chameleon players lose the game with a high probability for α-KL pairwise concealing strategies since the non-chameleons cannot statistically distinguish the chameleon at the voting stage.

Proposition 1. *For every α-KL pairwise concealing non-chameleon strategy π^{non}, there exists a chameleon strategy π^{ch} such that*

$$\Pr(\textit{Non-chameleons win}) \leq \frac{1}{N} + \frac{N-1}{N^2}\sqrt{\frac{(K-1)\alpha}{K}}.$$

For the trivial 0-KL pairwise concealing strategy that uses the same distribution for all words, the non-chameleons have a win rate of $(K-1)/NK$. We next consider revealing strategies.

Definition 2. *Let $L1(\mathcal{D}^{w_i}, \mathcal{D}^{w_j})$ be the L1-distance between distributions $\mathcal{D}^{w_i}$ and $\mathcal{D}^{w_j}$. A stationary non-chameleon strategy π^{non} is α-L1 pairwise revealing if*

$$L1(\mathcal{D}^{w_i}, \mathcal{D}^{w_j}) \geq \alpha \text{ for all } i \neq j \in [K].$$

Table 1: Numerical results from one hundred games of The Chameleon with players using nominal LLMs.

Non-ch. LLM	Chameleon: GPT-5		
	Non-ch. win ratio	Identification ratio	2nd-chance win ratio
GPT-5	0.00	0.64	1.00
GPT-4.1	0.03	0.45	0.93
GPT-4o	0.01	0.43	0.98
Gemini 2.5 Pro	0.06	0.35	0.83
Llama 3.1 70B	0.05	0.29	0.82
Qwen3 32B	0.02	0.23	0.91

We show that non-chameleons lose the game with high probability, as L1-pairwise revealing strategies lead to a correct guess of the secret word, and the chameleon wins the game in the second round even if it gets voted.

Proposition 2. *For every α-L1 pairwise revealing non-chameleon strategy π^{non} such that $\alpha \geq 1$, there exists a chameleon π^{ch} strategy such that*

$$\Pr(\textit{Non-chameleons win}) \leq 6(K-1)\left(\frac{2-\alpha}{\alpha}\right)^{\frac{(N-1)\alpha}{2\alpha-2}}.$$

A 2-L1 pairwise revealing strategy has unique responses for each secret word, leading to secret word inference with probability 1.

The Chameleon requires players to strategically reveal, conceal, and infer information. We use The Chameleon as an example to measure how suitable large language models are to environments where they interact with agents of unknown intentions.

3 EXPERIMENTS WITH LLM AGENTS

Setting. We instructed LLMs to play The Chameleon (with categories with 16 possible secret words) against each other. The games had four players. We varied the non-chameleon model, but used the same model for all three non-chameleon players. We report the results in Table 1 with GPT-5 as the chameleon, as it had the highest win as the chameleon.

Results. Our results show that non-chameleons (except for Qwen3) identify the chameleon with a probability higher than the trivially achievable level 25%. On the other hand, the identified chameleon has a second chance win ratio higher than 82% against all non-chameleon models. Consequently, the win ratios for all non-chameleon models range between 0% and 6%, while a strategy that outputs the same response for each secret word trivially achieves 23%. Combined with the theoretical analysis of revealing strategies, we deduce that non-chameleon LLMs reveal excessive information as they lose the game despite identifying the chameleon. Our results motivates the development of strategic LLMs by pointing to areas for improvement, such as the need to conceal information without deliberately misleading.

ACKNOWLEDGMENTS

This work was supported in part by the Army Research Office under Grant No. W911NF-23-1-0317 and the Office of Naval Research under Grant No. N00014-24-1-2432.

REFERENCES

[1] Sahar Abdelnabi, Amr Gomaa, Sarath Sivaprasad, Lea Schönherr, and Mario Fritz. 2024. Cooperation, competition, and maliciousness: LLM-stakeholders interactive negotiation. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

[2] Marc Alier, María José Casañ, and Daniel Amo Filvà. 2023. Smart Learning Applications: Leveraging LLMs for Contextualized and Ethical Educational Technology. In *International conference on technological ecosystems for enhancing multiculturality*. Springer, 190–199.

[3] Federico Bianchi, Patrick John Chia, Mert Yuksekgonul, Jacopo Tagliabue, Dan Jurafsky, and James Zou. 2024. How well can LLMs negotiate? NEGOTIATION-ARENA platform and analysis. In *Proceedings of the 41st International Conference on Machine Learning* (Vienna, Austria). Article 158, 17 pages.

[4] Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261* (2025).

[5] Yihong Dong, Xue Jiang, Zhi Jin, and Ge Li. 2024. Self-collaboration code generation via ChatGPT. *ACM Transactions on Software Engineering and Methodology* 33, 7 (2024), 1–38.

[6] Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783* (2024). Note: The used version is the instruction-tuned and quantized version hugging-quants/Meta-Llama-3.1-70B-Instruct-AWQ-INT4 from HuggingFace.

[7] Mert Karabacak and Konstantinos Margetis. 2023. Embracing large language models for medical applications: opportunities and challenges. *Cureus* 15, 5 (2023).

[8] Feng Lin, Dong Jae Kim, et al. 2024. When LLM-based code generation meets the software development process. *arXiv preprint arXiv:2403.15852* (2024).

[9] OpenAI. 2024. *GPT-4o*. https://openai.com/ Version: 2024-08-06.

[10] OpenAI. 2025. *GPT-4.1*. https://openai.com/ Version: 2025-04-14.

[11] OpenAI. 2025. *GPT-5*. https://openai.com/ Version: 2025-08-07.

[12] Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. 2023. Large language models encode clinical knowledge. *Nature* 620, 7972 (2023), 172–180.

[13] Rikki Tahta. 2017. The Chameleon Board Game. https://bigpotato.com/products/the-chameleon.

[14] Shijie Wu, Ozan Irsoy, Steven Lu, Vadim Dabravolski, Mark Dredze, Sebastian Gehrmann, Prabhanjan Kambadur, David Rosenberg, and Gideon Mann. 2023. Bloomberggpt: A large language model for finance. *arXiv preprint arXiv:2303.17564* (2023).

[15] Changrong Xiao, Sean Xin Xu, Kunpeng Zhang, Yufang Wang, and Lei Xia. 2023. Evaluating reading comprehension exercises generated by LLMs: A showcase of ChatGPT in education applications. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*. 610–625.

[16] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. 2025. Qwen3 Technical Report. arXiv:2505.09388 [cs.CL] https://arxiv.org/abs/2505.09388 Note: The used version is the instruction-tuned and quantized version Qwen/Qwen3-32B-AWQ from HuggingFace.

[17] Hongyang Yang, Xiao-Yang Liu, and Christina Dan Wang. 2023. Fingpt: Open-source financial large language models. *arXiv preprint arXiv:2306.06031* (2023).