Superficial Alignment and Subtle Divergence in LLM Decision-Making

Manuel Cherep*, **Nikhil Singh***, **Pattie Maes**
MIT
{mcherep,nsingh1,pattie}@mit.edu

Abstract

LLMs are being set loose in complex, real-world environments involving sequential decision-making and tool use. Often, this involves making choices on behalf of human users. Not much is known about the distribution of such choices, and how susceptible they are to different choice architectures. We perform a case study with a few such LLM models on a multi-attribute tabular decision-making problem, under the canonical default option nudge and additional prompting strategies. We show that, despite superficial similarities to human choice distributions, such models differ in subtle but important ways. First, they show much higher susceptibility to the default option nudge. Second, they diverge in points earned, being affected by factors like the idiosyncrasy of available prizes. Third, they diverge in information acquisition strategies: e.g. incurring substantial cost to reveal too much information, or selecting without revealing any. Finally, we show that simple prompt nudges like self-explanations can shift the choice distribution, and few-shot prompting with human data can induce greater alignment. These findings suggest that more information is needed before deploying models as agents or assistants acting on behalf of users in complex environments.

1 Introduction

We seem to want more from our language models than just a good conversation. Software agents powered by LLMs can now in principle browse the web [1], use spreadsheets [2], go shopping [3], make financial decisions [4], and make many kinds of choices while operating computer-based tools [5, 6]. Yet, we don't know how they choose. Do they choose what we would? Or do they systematically differ in important ways we should better understand before we hand the reins over? How easily and extensively can their choices be manipulated, intentionally or not? If simple nudges can significantly change such agents' decisions, they could have adverse effects on the people whose lives these decisions affect.

In this paper, we conduct a case study comparing LLM and human choices in a complex, sequential decision making task [7]. The task involves a meta-level decision making problem, wherein agents must make decisions about how to decide. The behavior of human agents in this task has been predicted using a resource rational model, and in particular how such human decision-making responds to nudges [7]. We construct a version of this task for LLMs, and examine how they make decisions and how this differs from their human participant counterparts. We also analyze how LLM decision-making is affected by a simple "default option" nudge, in which one option is labeled as a default. We show that, despite some superficial signs of alignment, LLM decisions depart substantially and unpredictably from human decision-making processes, and exhibit artifacts that reflect different meta-level strategies. Finally, we show some simple ways that we can further "nudge" models to better replicate human participants' decision making.

*Equal contribution.

NeurIPS 2024 Workshop on Behavioral Machine Learning.

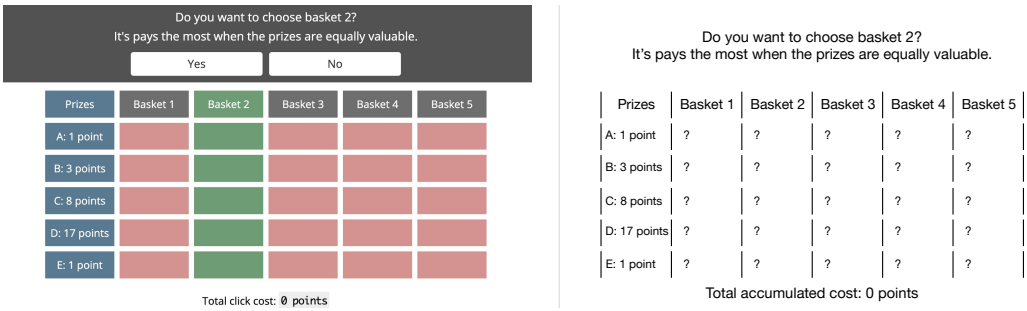

Figure 1: **(Left)** A screenshot from the original game setup [7] displaying a default option nudge, the number of prizes with their points, and the hidden cells for each basket. **(Right)** Our reconstruction of the game with just text and minor rephrasing, indicating hidden cells with a question mark.

2 Related Work

There's no guarantee that training LLMs with human-generated data leads to behavior that aligns with how we actually behave. Human behavior is complex, and often eludes our intuition. For example, the way people choose often contradicts traditional ideas about decision-making [8, 9], such as expected utility theory, which assume that people make rational choices. Instead, resource-rational models build on bounded rationality [10] to assume people choose rationally but are limited by their computational resources [11]. As another example, choice architecture (i.e. the variation of ways in which options are presented) influences how people choose [12]. Nudges (interventions on choice architecture) can be structured to alter people's choices in predictable ways [13], often towards beneficial outcomes, without limiting people's ability to make their own decisions. While these behaviors occur widely in people, the decision-making process in LLM-based agents is unknown and difficult to evaluate. Considering that these models are being used to simulate human behavior [14–18], it is important to study their implicit decision-making process.

Previous research studying behavior alignment has shown that LLMs model people as highly rational decision-makers [19], struggle to accurately model trade-offs seen in human behavior [20], exacerbate human biases [21], and show high variance in their performance as proxies for human behavior [18]. Moreover, LLMs are sensitive to small perturbations in prompts [22–25], the format of information like tabular data [26], the order of multiple-choice questions [27], and the prompt architecture [28, 23]; and are influenced by probabilities even in deterministic tasks [29]. Though people are deciding when and where to deploy these models, and could conceivably mitigate such issues by choosing responsibly, we are sometimes overconfident about the capabilities of LLMs [30]. Ultimately, better understanding how LLM-based agents make decisions, and how this differs from human decision-makers, might allow us to both design better choice architectures for LLMs, and make informed choices about when and how to deploy such agents.

3 Methods

To explore LLMs' implicit decision-making, we replicated an existing experimental paradigm for studying nudges with people [7]. The experiment consists of hidden decision-relevant information that agents can choose to reveal for a cost (see Figure 1). There are prizes with associated points p that apply equally to all different baskets with hidden cells B_i. The reward r for choosing a basket is $r = p \cdot B_i$. Revealing a cell costs 2 points, and 30 points $= \$0.01$ in reward. The goal is to choose the basket with the highest reward with minimal revealing cost. The process consists of a quiz, 2 practice rounds (unrewarded), and 32 scored test games. We ported the game to an LLM-compatible representation by (1) transforming the grid into a markdown tabular format, (2) providing callable functions to make decisions, and (3) adjusting instructions to match the textual format.

We also recreated the "default option" nudge, which is a simple, ubiquitous, and effective choice architecture that agents select unless they decline. Here, the basket with the most (unweighted) points is selected as the default; as such, it pays the most when prizes are equal. Half of the games are control (no nudge), and if the agent declines the nudge, the game continues like a control trial.

We sampled 10 participants from the original experiment to reconstruct the same game parameters. We conducted experiments with cutting-edge LLMs [31] (GPT-3.5-Turbo, GPT-4o-Mini, and GPT 4o; temperature $= 0.2$) with function calling capabilities for revealing, selecting, or accepting/declining a default option. Beyond testing them out-of-the-box, we attempted to "nudge" model behavior, to explore both the robustness of model decision-making to different choice architectures, and whether this can move it closer to the human behavior distribution. We tested four different conditions: (1) regular game, (2 & 3) simple self-explanation strategies, i.e. asking the model to provide an explanation *before* or *after* making a decision (hereafter, pre-response or post-response rationalization), and (4) leveraging few shot prompts [32–34] with 12 randomly sampled games (6 control, 3 from nudge-accepted and 3 from nudge-declined) from different trials, coming from unseen participants.

4 Results

All p-values in the results below are adjusted using the Benjamini-Hochberg correction.

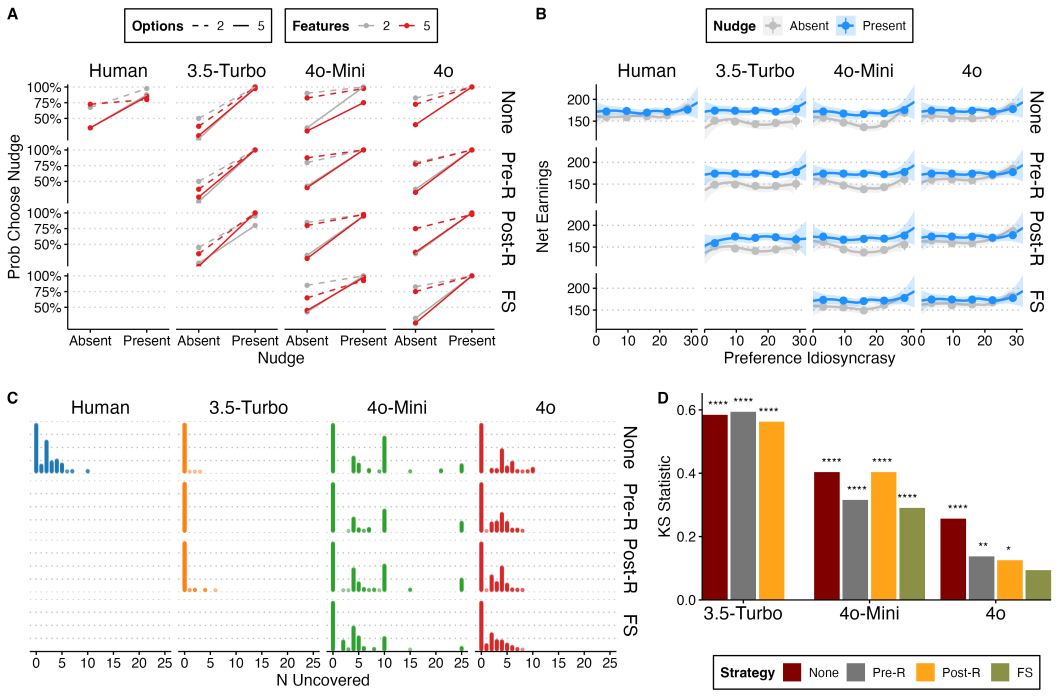

Figure 2: **(A)** Rate of choosing the "default" option in control vs. nudge trials. **(B)** Net earning points as a function of preference idiosyncrasy (L1 distance from the uniform weight vector) [7]. **(C)** Distribution of number of cells uncovered in a trial before a decision is made. **(D)** Kolmogorov–Smirnov test results comparing distributions in **C** to human participants (lower indicates better alignment). $*$ shows statistically significant difference ($* = p<0.05$; $** = p<0.01$; $**** = p<0.0001$). Strategies are "None" (regular game), "Pre-R" and "Post-R" (pre-response or post-response rationalization), and "FS" (few-shot) where we prompt with human game trials (except 3.5-Turbo due to context window limits), taken from unseen participants. **A+B** modeled after Callaway et al. [7].

4.1 Probability of Choosing the Default Option (Figure 2A)

At first glance, Figure 2A suggests alignment between human and model responses. The likelihood of choosing the default option appears similar overall, higher for less complex tasks, and increases with the nudge. However, a closer examination reveals considerable misalignment. We used a mixed-effects logistic regression model to predict the binary outcome of selecting the default option. The model incorporated data source (human vs. each LLM) and trial condition (control vs. nudge) as fixed effects, with random intercepts for choice set complexity factors (number of options and features), prompting method (e.g. explanations and few-shot prompting), and participant heterogeneity.

Without nudges, participants chose the default option with an estimated probability of 0.526 (95% CI [0.246, 0.791]). GPT-3.5-Turbo chose the default significantly less often, i.e. 0.281 (95% CI [0.106, 0.562]; odds ratio=0.352, p<0.0001). GPT-4 models were closer to participant behavior: GPT-4o-Mini at 0.61 (95% CI [0.324, 0.836]), and GPT-4o at 0.574 (95% CI [0.292, 0.815]). Neither GPT-4 model significantly differed from human participants (p>0.13 for both). With the nudge, human participants' probability rose to 0.899 (95% CI [0.714, 0.970]). Still, all models significantly surpassed this. GPT-3.5-Turbo increased to 0.983 (95% CI [0.940, 0.995]; odds ratio=6.550, p<0.0001), GPT-4o-Mini to 0.976 (95% CI [0.920, 0.993]; odds ratio=4.522, p<0.0001), and GPT-4o to 0.998 (95% CI [0.986, 1.000]; odds ratio=49.767, p<0.0001). While GPT-4o variants might approximate participant decision-making in the neutral context, they exhibit much higher sensitivity to the nudge.

4.2 Net Earnings (Figure 2B)

To examine earnings, we used a linear mixed-effects model predicting total (net) points earned, incorporating data source (human vs. each LLM), trial condition, and preference idiosyncrasy (L1 distance from the uniform weight vector) [7] as fixed effects, with the same random intercepts noted before. We used post-hoc marginal effect contrasts, marginalizing over preference idiosyncrasy.

Without the nudge, participants earned an estimated 162.89 points on average (95% CI [144.86, 180.92]). GPT-3.5-Turbo earned much less, at 146.98 points (95% CI [129.41, 164.54]; p<0.0001). GPT-4o-Mini also earned significantly less, with 149.86 points (95% CI [132.36, 167.37]; p<0.0001). GPT-4o's estimated average earnings of 164.63 points (95% CI [147.12, 182.13]) were higher but did not significantly differ from participants' (p=0.73). With the nudge present, participants' earnings increased to 171.81 points (95% CI [153.78, 189.83]). Interestingly, in this condition, none of the models significantly differed. GPT-3.5-Turbo earned 172.70 points (95% CI [155.13, 190.26]; p=0.83), GPT-4o-Mini earned 171.21 points (95% CI [153.70, 188.71]; p=0.61), and GPT-4o earned 173.98 points (95% CI [156.48, 191.49]; p=0.62). In contrast to the default option choices previously discussed, here the models show significantly more alignment with human decision-making *with* the nudge present. From Figure 2B, this seems especially evident as preference idiosyncrasy increases.

4.3 Strategies for Information Acquisition (Figure 2C+D)

To analyze information acquisition strategies, we examined the number of cells uncovered before making a choice. We used two sample Kolmogorov–Smirnov (KS) tests to compare the distributions of each model against human participants across different prompting methods.

Without specialized prompting ("None" condition), all models showed significant differences from human behavior. GPT-3.5-Turbo exhibited the largest deviation (D=0.584, p<0.0001), followed by GPT-4o-Mini (D=0.403, p<0.0001), and GPT-4o (D=0.256, p<0.0001). Pre-response and post-response rationalization ("Pre-R" and "Post-R" conditions) slightly shifted these differences. GPT-3.5-Turbo still differed the most (Pre-R: D=0.594, p<0.0001; Post-R: D=0.562, p<0.0001), followed by GPT-4o-Mini (Pre-R: D=0.316, p<0.0001; Post-R: D=0.403, p<0.0001), and GPT-4o (Pre-R: D=0.138, p=0.006; Post-R: D=0.125, p=0.015). In the few-shot prompting ("FS") condition, GPT-4o-Mini still showed a significant difference from human behavior (D=0.291, p<0.0001), but GPT-4o's distribution was not significantly different (D=0.0938, p=0.12). Figure 2C corroborates these findings: GPT-3.5-Turbo almost always answered immediately without uncovering cells. GPT-4o-Mini often uncovered *many* cells, incurring high costs, and tended to uncover in multiples of 5 suggesting simplistic strategies like uncovering entire rows or columns. GPT-4o's distribution appears most similar to human participants', suggesting that few-shot prompting (with different human data) might be an effective strategy to align decision-making, given a sufficiently strong base model.

5 Conclusion

This study compared LLM and human decision-making in a complex sequential reasoning task, examining effects of default options and strategies for eliciting human-like decision-making. Limitations include a small human sample, limited model and nudge variety, and prompt sensitivity. Using human example data proved helpful but may not always be available. Future research should expand to different settings, nudges, and models (like the recent o1 model, which we tried but can't yet call functions) to better understand LLM-based agent behavior in complex decision-making scenarios.

Acknowledgements

The project that gave rise to these results received the support of a fellowship from "la Caixa" Foundation (ID 100010434). The fellowship code is LCF/BQ/EU23/12010079. We thank Keyon Vafa for supportive comments.

References

[1] Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, et al. Webgpt: Browser-assisted question-answering with human feedback. *arXiv preprint arXiv:2112.09332*, 2021.

[2] Hongxin Li, Jingran Su, Yuntao Chen, Qing Li, and ZHAO-XIANG ZHANG. Sheetcopilot: Bringing software productivity to the next level through large language models. *Advances in Neural Information Processing Systems*, 36, 2024.

[3] Shunyu Yao, Howard Chen, John Yang, and Karthik Narasimhan. Webshop: Towards scalable real-world web interaction with grounded language agents. *Advances in Neural Information Processing Systems*, 35:20744–20757, 2022.

[4] Yangyang Yu, Haohang Li, Zhi Chen, Yuechen Jiang, Yang Li, Denghui Zhang, Rong Liu, Jordan W Suchow, and Khaldoun Khashanah. Finmem: A performance-enhanced llm trading agent with layered memory and character design. In *Proceedings of the AAAI Symposium Series*, volume 3, pages 595–597, 2024.

[5] Grégoire Mialon, Roberto Dessì, Maria Lomeli, Christoforos Nalmpantis, Ram Pasunuru, Roberta Raileanu, Baptiste Rozière, Timo Schick, Jane Dwivedi-Yu, Asli Celikyilmaz, et al. Augmented language models: a survey. *arXiv preprint arXiv:2302.07842*, 2023.

[6] Geunwoo Kim, Pierre Baldi, and Stephen McAleer. Language models can solve computer tasks. *Advances in Neural Information Processing Systems*, 36, 2024.

[7] Frederick Callaway, Mathew Hardy, and Thomas L Griffiths. Optimal nudging for cognitively bounded agents: A framework for modeling, predicting, and controlling the effects of choice architectures. *Psychological Review*, 2023.

[8] Amos Tversky and Daniel Kahneman. Judgment under uncertainty: Heuristics and biases: Biases in judgments reveal some heuristics of thinking under uncertainty. *science*, 185(4157):1124–1131, 1974.

[9] Daniel Kahneman and Amos Tversky. Prospect theory: An analysis of decision under risk. In *Handbook of the fundamentals of financial decision making: Part I*, pages 99–127. World Scientific, 2013.

[10] Herbert A Simon. A behavioral model of rational choice. *The quarterly journal of economics*, pages 99–118, 1955.

[11] Falk Lieder and Thomas L Griffiths. Resource-rational analysis: Understanding human cognition as the optimal use of limited computational resources. *Behavioral and brain sciences*, 43:e1, 2020.

[12] Richard H Thaler, Cass R Sunstein, and John P Balz. Choice architecture. *The behavioral foundations of public policy*, 2014.

[13] Richard H. Thaler and Cass Robert Sunstein. Nudge: Improving decisions about health, wealth, and happiness. 2008.

[14] Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th annual acm symposium on user interface software and technology*, pages 1–22, 2023.

[15] Joon Sung Park, Lindsay Popowski, Carrie Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. Social simulacra: Creating populated prototypes for social computing systems. In *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology*, pages 1–18, 2022.

[16] Lisa P Argyle, Ethan C Busby, Nancy Fulda, Joshua R Gubler, Christopher Rytting, and David Wingate. Out of one, many: Using language models to simulate human samples. *Political Analysis*, 31(3):337–351, 2023.

[17] Ryan Liu, Howard Yen, Raja Marjieh, Thomas L Griffiths, and Ranjay Krishna. Improving interpersonal communication by simulating audiences with language models. *arXiv preprint arXiv:2311.00687*, 2023.

[18] Tongshuang Wu, Haiyi Zhu, Maya Albayrak, Alexis Axon, Amanda Bertsch, Wenxing Deng, Ziqi Ding, Bill Guo, Sireesh Gururaja, Tzu-Sheng Kuo, et al. Llms as workers in human-computational algorithms? replicating crowdsourcing pipelines with llms. *arXiv preprint arXiv:2307.10168*, 2023.

[19] Ryan Liu, Jiayi Geng, Joshua C Peterson, Ilia Sucholutsky, and Thomas L Griffiths. Large language models assume people are more rational than we really are. *arXiv preprint arXiv:2406.17055*, 2024.

[20] Ryan Liu, Theodore R Sumers, Ishita Dasgupta, and Thomas L Griffiths. How do large language models navigate conflicts between honesty and helpfulness? *arXiv preprint arXiv:2402.07282*, 2024.

[21] Katherine Van Koevering and Jon Kleinberg. How random is random? evaluating the randomness and humaness of llms' coin flips. *arXiv preprint arXiv:2406.00092*, 2024.

[22] Jiongxiao Wang, Zichen Liu, Keun Hee Park, Zhuojun Jiang, Zhaoheng Zheng, Zhuofeng Wu, Muhao Chen, and Chaowei Xiao. Adversarial demonstration attacks on large language models. *arXiv preprint arXiv:2305.14950*, 2023.

[23] Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. Calibrate before use: Improving few-shot performance of language models. In *International conference on machine learning*, pages 12697–12706. PMLR, 2021.

[24] Kaijie Zhu, Jindong Wang, Jiaheng Zhou, Zichen Wang, Hao Chen, Yidong Wang, Linyi Yang, Wei Ye, Yue Zhang, Neil Zhenqiang Gong, et al. Promptbench: Towards evaluating the robustness of large language models on adversarial prompts. *arXiv preprint arXiv:2306.04528*, 2023.

[25] Melanie Sclar, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. 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*, 2023.

[26] Yuan Sui, Mengyu Zhou, Mingjie Zhou, Shi Han, and Dongmei Zhang. Table meets llm: Can large language models understand structured table data? a benchmark and empirical study. In *Proceedings of the 17th ACM International Conference on Web Search and Data Mining*, pages 645–654, 2024.

[27] Pouya Pezeshkpour and Estevam Hruschka. Large language models sensitivity to the order of options in multiple-choice questions. *arXiv preprint arXiv:2308.11483*, 2023.

[28] Melanie Brucks and Olivier Toubia. Prompt architecture can induce methodological artifacts in large language models. *Available at SSRN 4484416*, 2023.

[29] R Thomas McCoy, Shunyu Yao, Dan Friedman, Matthew Hardy, and Thomas L Griffiths. Embers of autoregression: Understanding large language models through the problem they are trained to solve. *arXiv preprint arXiv:2309.13638*, 2023.

[30] Keyon Vafa, Ashesh Rambachan, and Sendhil Mullainathan. Do large language models perform the way people expect? measuring the human generalization function. *arXiv preprint arXiv:2406.01382*, 2024.

[31] OpenAI. Openai api - models documentation, 2024. Accessed: 2024-09-13.

[32] Tom B Brown. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 2020.

[33] Sanyam Kapoor, Nate Gruver, Manley Roberts, Katherine Collins, Arka Pal, Umang Bhatt, Adrian Weller, Samuel Dooley, Micah Goldblum, and Andrew Gordon Wilson. Large language models must be taught to know what they don't know. *arXiv preprint arXiv:2406.08391*, 2024.

[34] Omar Shaikh, Michelle Lam, Joey Hejna, Yijia Shao, Michael Bernstein, and Diyi Yang. Show, don't tell: Aligning language models with demonstrated feedback. *arXiv preprint arXiv:2406.00888*, 2024.