Evaluating LLMs as Participant Simulators for Behavioral Science

Sabrina Namazova and **Alessandra Brondetta** *Osnabrück University*
{snamazova,alessandra.brondetta}@uni-osnabrueck.de
Younes Strittmatter *Princeton University* ystrittmatter@princeton.edu
Matthew Nassar *Brown University* matthew_nassar@brown.edu
Sebastian Musslick *Osnabrück University, Brown University*
sebastian.musslick@uos.de

Abstract

Collecting data from human participants in cognitive experiments is a costly and time-consuming aspect of behavioral science. One promising direction is to fine-tune large language models (LLMs) on human behavior to act as participant simulators. In this work, we outline key criteria such simulators must satisfy and evaluate how well state-of-the-art fine-tuned LLMs meet them. Our analyses indicate that although such LLMs achieve high predictive accuracy, their generative behavior—a key requirement for simulating participants—systematically diverges from human data. To probe this discrepancy, we examine the role of experimental information provided to the model. The results replicate prior work showing that LLMs primarily act as autoregressive predictors: they excel at forecasting responses from prior behavior but fail to integrate information about the experimental context, leading to weak generative performance. These findings highlight both the potential and methodological challenges of using LLMs as synthetic participants, emphasizing the need for careful validation before integrating them into behavioral research.

Keywords: automated scientific discovery | experimental design | automated cognitive science | cognitive psychology

1 Introduction

Simulators have revolutionized scientific practice across the natural sciences. By generating data that reliably approximate real-world phenomena, they enable scientists to accelerate hypothesis testing and optimize experimental designs [Jumper et al., 2021, Krenn et al., 2016]. This is perhaps best illustrated by AlphaFold, a Nobel-prize winning simulator in chemistry that predicts protein structures from amino acid sequences, enabling rapid prototyping of molecular interactions, drug targets, and protein functions [Jumper et al., 2021]. In the behavioral sciences, a reliable participant simulator—a system capable of producing human-like behavior across cognitive tasks—would represent a similarly transformative advance, allowing for fast and cheap in-silico hypothesis testing before slow and expensive in-vivo validation [Musslick et al., 2024].

In this study, we assess how well LLMs fine-tuned to human behavior meet a core criterion of participant simulators: the ability to generate human-like behavior from scratch. We find that although fine-tuned LLMs show strong predictive accuracy, their generative behavior diverges from human data. We hypothesize that this divergence reflects their tendency to condition their predictions about future behavior primarily on past behavior instead of task-related information.

2 Related Work

Recent work has explored using LLMs as simulators of human behavior [Hardy et al., 2023, Manning et al., 2024, Strittmatter and Musslick, 2025, Binz et al., 2023], for example in automated cognitive science [Musslick et al., 2024, 2025]. LLMs can reproduce human-like patterns in psychological tasks [Binz and Schulz, 2023] and, when fine-tuned to predict trial-by-trial human responses, even outperform domain-specific cognitive models in predictive performance [Binz et al., 2025, Zhu et al., 2025]. However, analyses of Centaur—an LLM fine-tuned to predict human behavior in cognitive experiments—suggest that its advantage often comes from exploiting choice history rather than task structure: the model outperforms cognitive baselines even without task descriptions, but underperforms when deprived of behavioral history [Xie and Zhu, 2025].

Importantly, predictive and generative performance can diverge. As Palminteri et al. [2017] note, computational models may predict future behavior accurately from past data yet fail to generate plausible behavior from scratch. This distinction is central to computational cognitive modeling and motivates our investigation: comparing the predictive and generative performance of Centaur with its base LLM and domain-specific cognitive models.

3 Methods

To assess Centaur's capacity as a participant simulator, we evaluated its predictive performance and generative performance across three cognitive tasks, comparing it to Llama-Instruct and domain-specific cognitive models. Predictive performance is defined as the likelihood that the model's choices matches the human choices given prior human choices and task information. Specifically, we computed the average negative log-likelihood of observed choices given the models' predictions. Generative performance is defined as the extent to which the model reproduces human-like behavioral patterns when simulated from scratch on the task. In this simulation, the models' own past choices were appended trial by trial, without anchoring it on prior human choices.

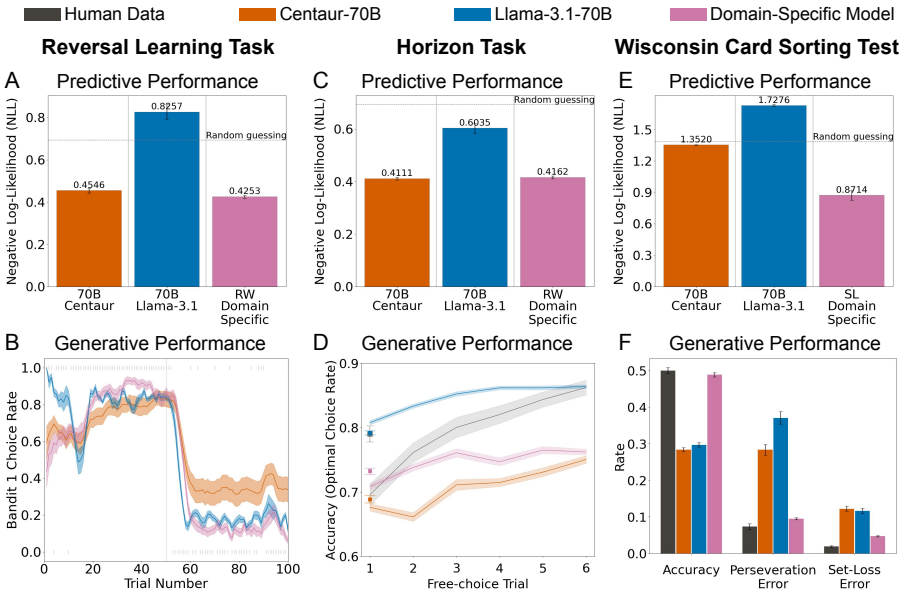

Figure 1: Predictive and generative performance of Centaur, its base model (Llama 3.1), and domain-specific models: Rescorla–Wagner model (RW) [Binz et al., 2025], and sequential learning model (SL) [Bishara et al., 2010]. Models were tested in a reversal learning task (A–B), horizon-dependent bandit task (C–D), and Wisconsin Card Sorting Test (E–F). Predictive performance was assessed in terms of negative log-likelihood (NLL; lower values indicate better fits). Generative performance shows: (B) choice rate for bandit 1 (rewarded 80% until trial 50, then reversed); (D) proportion of optimal choices by horizon length (1 vs. 6 free choices); and (F) accuracy and error type distributions in the card-sorting task. Error bars/shaded areas show standard error of the mean across participants or simulated instances.

As the first cognitive task, we considered an adaptation of a two-armed reversal learning paradigm on which Centaur was trained [Eckstein et al., 2022]. On every trial in this task, participants choose between two bandits: one yielding an 80% reward probability and the other 20%. The reward contingencies reverse in the middle of the experiment (trial 50), requiring participants to adapt their choices. Humans and animals typically reverse their choices several trials after the reward reveral [Eckstein et al., 2022, Izquierdo et al., 2017]. Therefore, we examined whether Centaur exhibited a similar delayed reversal pattern in its generative behavior. As the second cognitive task, we considered the horizon-dependent bandit paradigm, which was also included in Centaur's training set. In this task,participants again chose between two bandits but faced different time horizons—either 1 or 6 free choices—requiring them to balance exploration and exploitation. Humans and animals tend to forego the highest expected-value action more often in longer horizons to explore options they can later exploit

[Wilson et al., 2014]. Thus, we examined whether the models exhibited a similar horizon-dependent exploration pattern. Finally, we evaluated Centaur on the Wisconsin Card Sorting Test (WCST) [Steinke et al., 2020], which was not included in its training set. In this task, participants must infer and apply an unstated card-sorting rule (e.g., based on color, shape, or number of objects) from feedback, and flexibly switch when the rule changes unexpectedly. Humans generally perform well on this task but commit two characteristic types of errors: perseveration errors (failure to adapt to a new rule) and set-loss errors (failure to maintain the current rule). We therefore examined accuracy and error types to assess the similarity of Centaur's performance to that of human participants.

We evaluated each model's performance against ground-truth data generated by a Rescorla–Wagner agent in a reversal learning task, as well as against behavioral data from human participants in the horizon task [Wilson et al., 2014] and the Wisconsin Card Sorting Test [Steinke et al., 2020]. In addition, we analyzed Centaur's and Llama's performance in the reversal learning task under two prompting conditions: (a) a baseline condition, in which the model received task instructions, choice history, and reward feedback; and (b) a partial-feedback condition, in which the model received instructions and choice history but no reward feedback.

4 Results and Discussion

Our findings suggest that while Centaur achieves high predictive performance on tasks it was trained on, it still struggles to reproduce human-like behavior in those tasks, including the qualitative hallmarks of behavior that the tasks themselves were designed to measure (i.e., reversals, horizon effects; Figure 1). Moreover, on the task outside its fine-tuning set, it performed worse than the domain-specific model.

Strikingly, removing task-relevant information had little effect on Centaur's predictive accuracy (Figure 2); the model continued to outperform a standard cognitive model (Rescorla–Wagner) in the reversal learning task. This result aligns with previous findings suggesting that Centaur may rely heavily on participants' past choices to predict behavior [Xie and Zhu, 2025], achieving high predictive accuracy even in absence of task-related information. However, when required to perform a task from scratch, the model's failure to incorporate task-relevant information leads to poor generative performance.

By contrast, removing task-relevant information had a stronger effect on base LLM Llama. Curiously, removing this information (partial feedback) improved Llama's predictive performance. This counterintuitive improvement arose because full task information led to overconfident choices (probabilities spiking near 0 or 1), whereas partial feedback scattered its predictions, introducing uncertainty that better aligned with human choice patterns overall (Figure 2C).

Taken together, these results point to a potential limitation of fine-tuning LLMs to predict human behavioral data: while fine-tuning enhances predictive alignment with observed human data, it may achieve such alignment by learning

autoregressive biases (i.e., overrely on past choices while ignoreing task-relevant information). The latter is known to hamper generative performance [Palminteri et al., 2017].

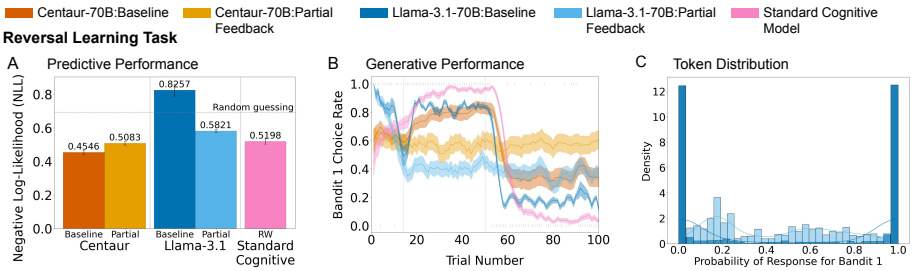

Figure 2: Predictive and generative performance in the reversal learning task under different prompt conditions. (A) Predictive performance in terms of negative log-likelihood (lower values indicate better fits). (B) Generative performance shown as the proportion of choosing bandit 1 (80% vs. 20% reward) across trials, with reward probabilities reversed at trial 50. (C) Token distribution for the base model LLama-3.1-70B.

5 Conclusion

While Centaur, an LLM fine-tuned to human behavior, achieves strong predictive accuracy, its generative behavior diverges systematically from human data. Consistent with prior work [Xie and Zhu, 2025], our results suggest that conventional fine-tuning may bias LLMs to overrely on past behavior rather than task information, mirroring the autoregressive nature of human behavior in cognitive tasks. Approaches such as weighted fine-tuning, which emphasizes trials diagnostic of task-dependent behavioral signatures, may help steer LLMs toward more reliable participant simulators, bringing us closer to a transformative tool for AI-driven discovery in the behavioral sciences.

6 Data and Code Availability

Data and code for all reported analyses all figures are available at the following repository: github.com/sabrinaholmes/centaur_eval_simulator.

References

M. Binz and E. Schulz. Using cognitive psychology to understand gpt-3. *Proceedings of the National Academy of Sciences*, 120(6):e2218523120, 2023.

M. Binz, S. Alaniz, A. Roskies, B. Aczel, C. T. Bergstrom, C. Allen, D. Schad, D. Wulff, J. D. West, and Q. Zhang. How should the advent of large language models affect the practice of science? *arXiv preprint arXiv:2312.03759*, 2023.

M. Binz, E. Akata, M. Bethge, F. Brändle, F. Callaway, J. Coda-Forno, P. Dayan, C. Demircan, M. K. Eckstein, N. Éltető, et al. A foundation model to predict and capture human cognition. *Nature*, pages 1–8, 2025.

A. J. Bishara, J. K. Kruschke, J. C. Stout, A. Bechara, D. P. McCabe, and J. R. Busemeyer. Sequential learning models for the wisconsin card sort task: Assessing processes in substance dependent individuals. *Journal of Mathematical Psychology*, 54(1):5–13, 2010.

M. K. Eckstein, S. L. Master, R. E. Dahl, L. Wilbrecht, and A. G. Collins. Reinforcement learning and bayesian inference provide complementary models for the unique advantage of adolescents in stochastic reversal. *Developmental Cognitive Neuroscience*, 55:101106, 2022.

M. Hardy, I. Sucholutsky, B. Thompson, and T. Griffiths. Large language models meet cognitive science: Llms as tools, models, and participants. In *Proceedings of the annual meeting of the cognitive science society*, volume 45, 2023.

A. Izquierdo, J. L. Brigman, A. K. Radke, P. H. Rudebeck, and A. Holmes. The neural basis of reversal learning: An updated perspective. *Neuroscience*, 345: 12–26, 2017.

J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek, and A. Potapenko. Highly accurate protein structure prediction with AlphaFold. *Nature*, 596(7873):583–589, 2021.

M. Krenn, M. Malik, R. Fickler, R. Lapkiewicz, and A. Zeilinger. Automated search for new quantum experiments. *Physical Review Letters*, 116(9):090405, 2016.

B. S. Manning, K. Zhu, and J. J. Horton. Automated social science: Language models as scientist and subjects. Technical report, National Bureau of Economic Research, 2024.

S. Musslick, Y. Strittmatter, and M. Dubova. Closed-loop scientific discovery in the behavioral sciences. PsyArXiv, 2024. Preprint.

S. Musslick, L. K. Bartlett, S. H. Chandramouli, M. Dubova, F. Gobet, T. L. Griffiths, J. Hullman, R. D. King, J. N. Kutz, C. G. Lucas, et al. Automating the practice of science: Opportunities, challenges, and implications. *Proceedings of the National Academy of Sciences*, 122(5):e2401238121, 2025.

S. Palminteri, V. Wyart, and E. Koechlin. The importance of falsification in computational cognitive modeling. *Trends in Cognitive Sciences*, 21(6):425–433, 2017.

A. Steinke, F. Lange, and B. Kopp. Parallel model-based and model-free reinforcement learning for card sorting performance. *Scientific Reports*, 10(1): 15464, 2020.

Y. Strittmatter and S. Musslick. Sweetbean: A declarative language for behavioral experiments with human and artificial participants. *Journal of Open Source Software*, 10(107):7703, 2025.

R. C. Wilson, A. Geana, J. M. White, E. A. Ludvig, and J. D. Cohen. Humans use directed and random exploration to solve the explore–exploit dilemma. *Journal of Experimental Psychology: General*, 143(6):2074, 2014.

H. Xie and J.-Q. Zhu. Centaur may have learned a shortcut that explains away psychological tasks. PsyArXiv, 2025. Preprint.

J.-Q. Zhu, H. Xie, D. Arumugam, R. C. Wilson, and T. L. Griffiths. Using reinforcement learning to train large language models to explain human decisions. *arXiv preprint arXiv:2505.11614*, 2025.