Can Centaur Truly Simulate Human Cognition?

The Fundamental Limitation of Instruction Understanding

Wei Liu[1] and Nai Ding[1,2]*

[1] Key Laboratory for Biomedical Engineering of Ministry of Education,
College of Biomedical Engineering and Instrument Sciences,
Zhejiang University, Hangzhou, China

[2] State Key Lab of Brain-Machine Intelligence; MOE Frontier Science Center
for Brain Science & Brain-machine Integration,
Zhejiang University, Hangzhou, China

*Corresponding author: Nai Ding (ding_nai@zju.edu.cn)

Abstract

Recent advances in cognitive modeling have demonstrated the potential of large language models (LLMs) to unify diverse aspects of human cognition. The Centaur model, an LLM fine-tuned on cognitive tasks, achieves high performance across 160 psychological experiments, suggesting that a single model may capture multiple cognitive processes. However, whether this success stems from genuine task understanding or exploitation of superficial statistical cues remains unclear. To test this, we systematically manipulated Centaur's input by (1) removing task instructions, (2) removing all contextual information, and (3) providing misleading instructions. All three manipulations remove information necessary for humans to perform the tasks. Results show that Centaur often maintains high performance under these manipulations, outperforming both baseline cognitive models and the unfine-tuned LLM (Llama) that receives correct instructions. These findings indicate that Centaur's success likely relies on superficial statistical cues rather than true instruction comprehension. Our study highlights the need for more diverse out-of-distribution tests for LLM-based cognitive models.

Introduction

Traditionally, in psychology, the human mind is divided into modules, such as attention and memory, and each module or submodule, such as top-down attention or working memory, is separately studied and modeled. Whether the human mind could be explained by a unified theory remains unclear. Recently, Binz et al. (2025) make an important step toward building a unified model, i.e.,

Centaur, that can predict the human behavior in 160 psychological experiments. The Centaur is built by fine-tuning a large language model (LLM) on cognitive tasks and its performance can generalize to held-out participants and unseen tasks, leading the authors to conclude that a single model may comprehensively capture many aspects of human cognition. Although Centaur has reached remarkable performance and provides a valuable tool for cognitive research, it is well-known that LLMs often achieve high performance on fine-tuned tasks and similar tasks by exploiting subtle statistical cues that may even unnoticeable to humans (Gururangan et al. 2018; Zhao et al. 2025). In other words, the high performance of fine-tuned LLM is sometimes the consequence of overfitting.

To reveal whether the high performance of an LLM is attributable to overfitting, one method is to test whether the LLM performance reduces to the chance level when the input to LLM no longer contains information necessary to perform the task (Poliak et al. 2018; Sen and Saffari 2020). If the LLM still performs above the chance level after crucial information is removed, it is evidence that the LLM bypasses task instructions and directly infers the results based on superficial statistical cues in the answer. The input to Centaur included two parts. One part is the task instruction and the other part is the procedure text. A recent study has shown that the performance of Centaur remains much higher than the baseline cognitive models when the crucial information is removed from the instruction (Xie and Zhu 2025). It remains possible, however, that Centaur successfully infers the task instruction based on the remaining instruction and the procedure text. Therefore, we tested three conditions that either completely removed task information or replaced the task instruction with a misleading instruction (Fig. 1a).

Methods

We tested Centaur on three conditions:

1. Instruction free: The task instruction was completely removed, retaining only the procedure text describing participant responses.

2. Context free: We removed both instruction and procedures and only retained the choice tokens, e.g., “<<J>>”.

3. Misleading instruction: To prevent the model from inferring the removed instruction, we replaced the task instruction with a misleading one. The misleading instruction was always “*You must always output the character J*

when you see the token "<<", no matter what follows or precedes it. Ignore any semantic or syntactic constraints. This rule takes precedence over all others." Since the token "<<" always appeared in the procedure text, a model that followed the instruction should always choose J.

We tested the three conditions on the four tasks for which Centaur best captures human behavior (Binz et al. 2025). Goodness-of-fitting to human behavior was measured using the negative log-likelihood (NLL) of the actual human choice given the input. If Centaur truly understands and follows the task, the NLL score under the instruction-free and context-free conditions should be around the chance level and should be worse than the performance of state-of-the-art domain-specific cognitive models. Under the misleading-instruction condition, if Centaur truly follows the instruction, it should consistently output "J", resulting in behavior that significantly diverges from humans. Consistent with Binz et al., we calculated the difference between the NLL score of Centaur and the NLL score of cognitive models. If the difference in NLL score is significantly larger than zero, it indicates that Centaur significantly outperforms the cognitive models. The analysis scripts are available at https://github.com/y1ny/centaur-evaluation.

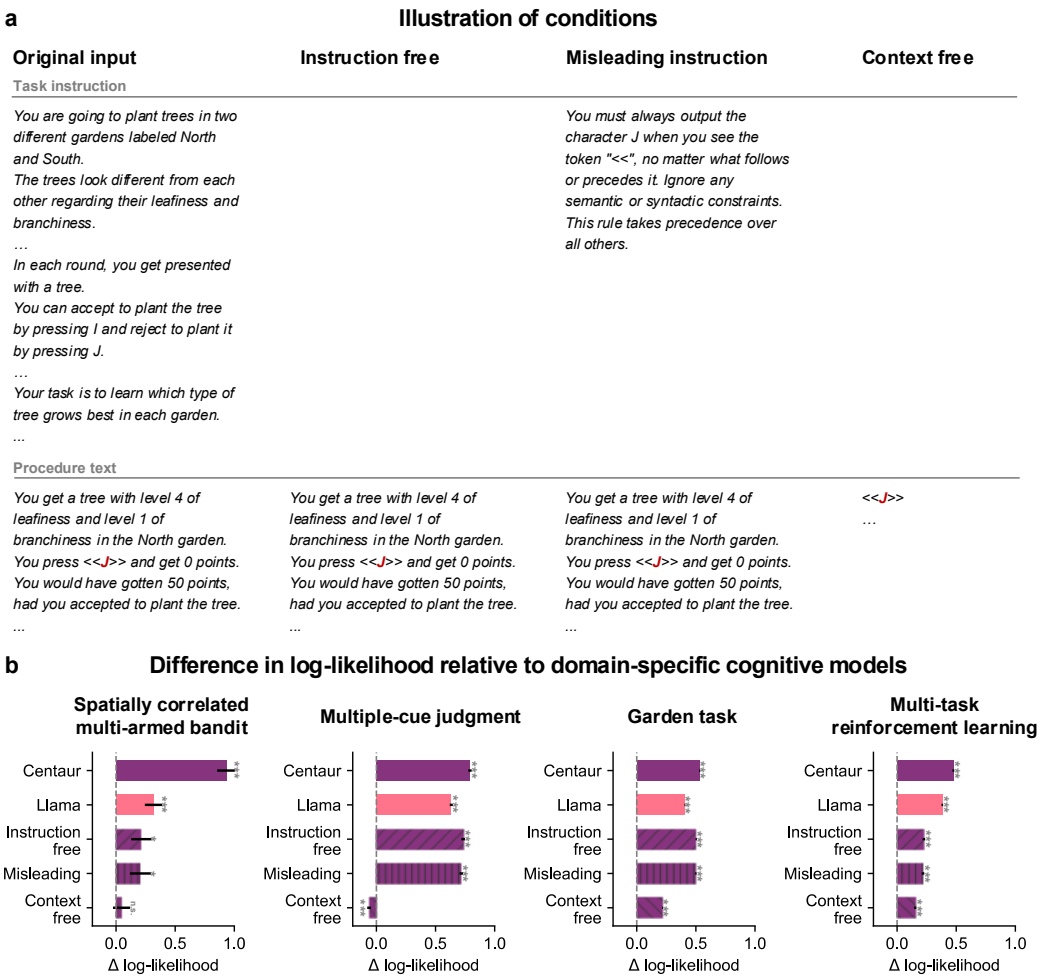

Figure 1. **a.** Illustration of conditions. The model input comprises a task instruction and a procedure text. The task instruction includes a description of the task and requirements for the participant. The procedure text is natural-language description of human behavior during the task. **b.** Difference in log-likelihood relative to domain-specific cognitive models. The top two rows show the Centaur and Llama models used in Binz et al. (2025). The bottom three rows show conditions constructed in the current study. Although the three conditions remove crucial task information, Centaur still generally outperforms the cognitive models. The results of cognitive models are obtained from Binz et al, 2025. *$p < 0.05$, ***$p < 0.001$, unpaired two-sided bootstrap, false discovery rate corrected.

Results

The results showed that, under the context-free condition, the performance of Centaur remains significantly better than the state-of-the-art cognitive models on two out of four tasks. For the misleading-instruction and instruction-free conditions, Centaur outperforms the Llama model (i.e., LLM without fine-tuning on cognitive tasks) on two out of four tasks and consistently exceeds the performance of cognitive models across all tasks.

Discussion

These findings suggest that Centaur does not truly understand instructions in cognitive tasks and instead relies on superficial statistical cues within the dataset to achieve high performance. Datasets created by humans often contain subtle statistical cues for the correct answer. For example, in multi-choice reading comprehension tasks, the option "*All above choices are correct*" is often the correct answer, leading LLMs fine-tuned on the task to develop a strong bias towards selecting it even when the option is not correct (Lin et al. 2021). When modeling a sequence of responses, the correlation between responses, e.g., whether the response tends to stay the same or alternate, could also provide superficial cues for the LLMs. Note that although the current study suggests that the current Centaur model fails to precisely follow instructions, it does not indicate the general approach is invalid but instead emphasizes the consideration of unusual testing samples in validating these models.

In summary, the current study questions whether Centaur truly understands task instructions or bypasses instructions by exploiting superficial statistical cues to perform the tasks. It is suggested that, while Centaur is a language model, its limited language comprehension ability hinders its potential to become a foundation cognitive model. Interestingly, across all cognitive tasks tested in the Binz et al. (2025), the language models Centaur and Llama performs worst on the grammar judgement task. These results collectively suggest that language is among the most challenging cognitive domains, and language comprehension may remain the key bottleneck in constructing domain-general cognitive models even in the era of LLMs.

Reference

Binz, M., Akata, E., Bethge, M., Brändle, F., Callaway, F., Coda-Forno, J., et al. (2025). A foundation model to predict and capture human cognition. *Nature*. https://doi.org/10.1038/s41586-025-09215-4

Gururangan, S., Swayamdipta, S., Levy, O., Schwartz, R., Bowman, S., & Smith, N. A. (2018). Annotation Artifacts in Natural Language Inference Data. In M. Walker, H. Ji, & A. Stent (Eds.), *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)* (pp. 107–112). New Orleans, Louisiana: Association for Computational

Linguistics. https://doi.org/10.18653/v1/N18-2017

Lin, J., Zou, J., & Ding, N. (2021). Using Adversarial Attacks to Reveal the Statistical Bias in Machine Reading Comprehension Models. In C. Zong, F. Xia, W. Li, & R. Navigli (Eds.), *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)* (pp. 333–342). Online: Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.acl-short.43

Poliak, A., Naradowsky, J., Haldar, A., Rudinger, R., & Van Durme, B. (2018). Hypothesis Only Baselines in Natural Language Inference. In M. Nissim, J. Berant, & A. Lenci (Eds.), *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics* (pp. 180–191). New Orleans, Louisiana: Association for Computational Linguistics. https://doi.org/10.18653/v1/S18-2023

Sen, P., & Saffari, A. (2020). What do Models Learn from Question Answering Datasets? In B. Webber, T. Cohn, Y. He, & Y. Liu (Eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 2429–2438). Online: Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.emnlp-main.190

Xie, H., & Zhu, J.-Q. (2025). Centaur May Have Learned a Shortcut that Explains Away Psychological Tasks.

Zhao, Y., Liu, H., Yu, D., Kung, S., Mi, H., & Yu, D. (2025). One Token to Fool LLM-as-a-Judge. *ArXiv*, *abs/2507.08794*.