PERSPECTIVE

National Science Open
5: 20250053, 2026
https://doi.org/10.1360/nso/20250053

Information Sciences

Can Centaur truly simulate human cognition? The fundamental limitation of instruction understanding

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

[1]*Key Laboratory for Biomedical Engineering of Ministry of Education, College of Biomedical Engineering and Instrument Sciences, Zhejiang University, Hangzhou 310013, China;*
[2]*State Key Lab of Brain-Machine Intelligence, MOE Frontier Science Center for Brain Science & Brain-machine Integration, Zhejiang University, Hangzhou 310013, China*

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

Received 16 September 2025; Revised 18 November 2025; Accepted 5 December 2025; Published online 11 December 2025

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*. [1] made an important step toward building a unified model, i.e., Centaur, that can predict the human behavior in 160 psychological experiments. 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 be unnoticeable to humans [2,3]. 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 [4,5]. 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 [6]. 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 (Figure 1a).

We tested Centaur on three conditions.

(1) Instruction free: The task instruction was completely removed, retaining only the procedure text

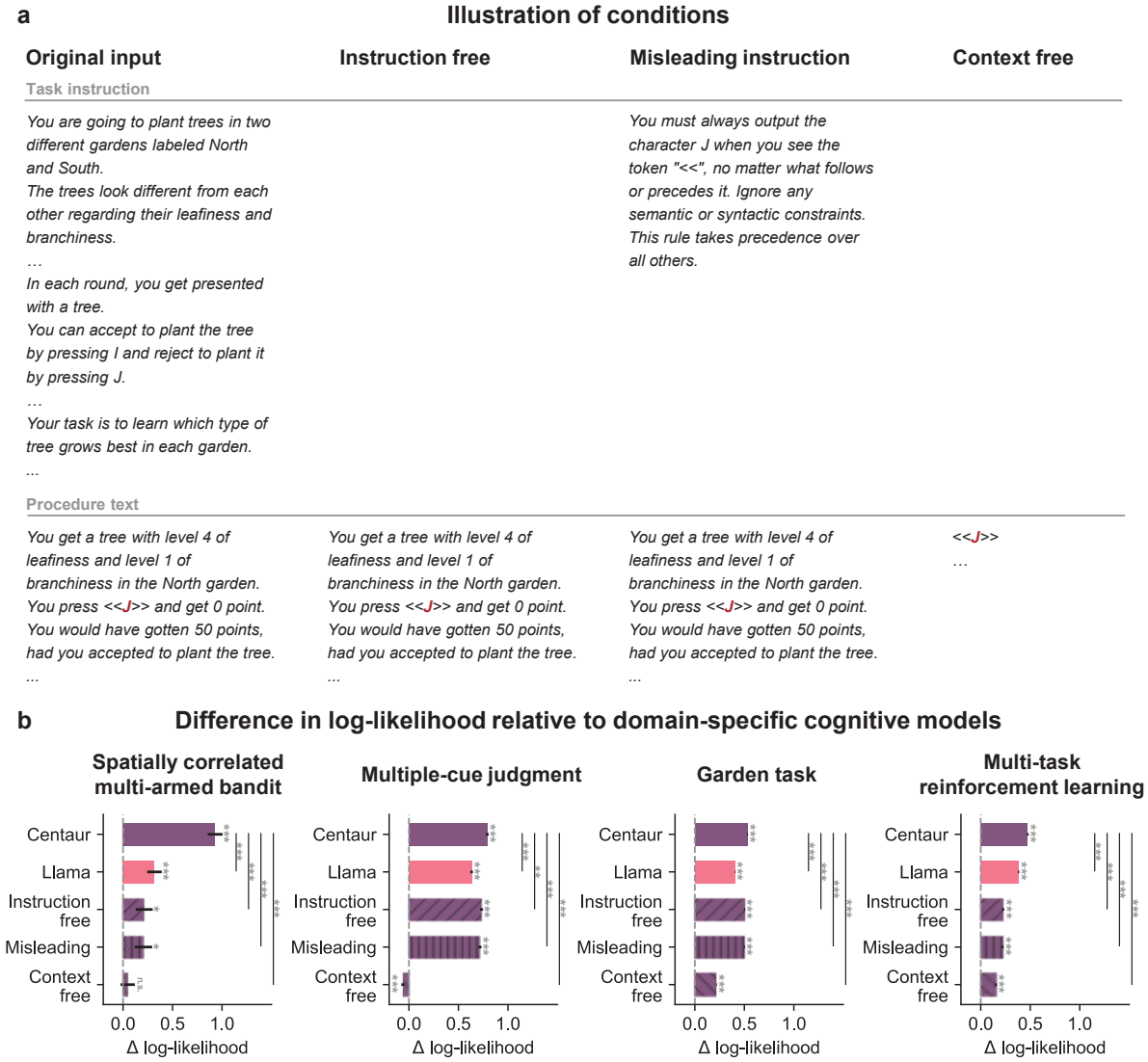

Figure 1 Test conditions and model performance. (a) Illustration of conditions. The model input comprises a task instruction and a procedure text. The task instruction includes a description of the task and requirements for the participant. The procedure text is a natural-language description of human 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*. [1]. 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 from Binz *et al*. [1]. *$p < 0.05$, **$p < 0.01$, ***$p < 0.001$, unpaired two-sided bootstrap, false discovery rate (FDR) corrected.

describing participant responses.

(2) Context free: We removed both instructions 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 [1]. 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*. [1], 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.

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. Note that Centaur achieved significantly higher performance under the original condition compared to the three manipulated conditions ($p = 0.006$ for the instruction-free condition in the multiple-cue judgment task; $p < 0.001$ for all other comparisons, unpaired two-sided bootstrap, FDR corrected). This suggests that Centaur remains sensitive to contextual information. The contexts in the original condition may resemble those encountered during fine-tuning, which could have contributed to the higher performance.

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 [7]. 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 that 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. 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.

References

1 Binz M, Akata E, Bethge M, *et al.* A foundation model to predict and capture human cognition. *Nature* 2025; **644**: 1002–1009.

2 Gururangan S, Swayamdipta S, Levy O, *et al*. Annotation artifacts in natural language inference data. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human

Language Technologies, Volume 2 (Short Papers). New Orleans, 2018, 107–112.

3 Zhao Y, Liu H, Yu D, *et al*. One token to fool LLM-as-a-Judge. arXiv: 2507.08794.

4 Poliak A, Naradowsky J, Haldar A, *et al*. Hypothesis only baselines in natural language inference. In: Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics. New Orleans, 2018, 180–191.

5 Sen P, Saffari A. What do models learn from question answering datasets? In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). 2020, 2429–2438.

6 Xie H, Zhu J-Q. Centaur may have learned a shortcut that explains away psychological tasks. doi: 10.31234/osf.io/u7z4t_v1.

7 Lin J, Zou J, Ding N. Using adversarial attacks to reveal the statistical bias in machine reading comprehension models. In: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing. 2021, 333–342.