

Accurately assessing the thinking capabilities of AI.

New research suggests that AI may not truly understand like humans. The Centaur model is suspected of merely memorizing data instead of simulating real thinking.

A new wave of AI research is returning to address one of psychology's oldest questions: whether the human mind can be explained by a single, unified theory.

For decades, psychologists have debated whether processes like memory, attention, and decision-making belong to a single system or should be studied separately. Now, the development of artificial intelligence is offering a new approach to examining this question — and also raising many questions about how well AI truly 'understands'.

The Centaur AI model and its ambition to mimic human thinking.

In July 2025, a study published in the journal *Nature* introduced an AI model called **Centaur**. This system is built upon existing large-scale language models, then refined using data from psychological experiments.

Centaur's goal is to simulate how humans think and make decisions. According to the research team, the model can replicate human-like responses in approximately 160 different cognitive tasks, including behavioral control and decision-making.

This finding quickly attracted the attention of researchers, as it suggested that AI could be getting closer to building a general cognitive model for humans — something that psychology has been searching for for decades.

However, a new study published in *National Science Open* comes to the opposite conclusion. Researchers from Zhejiang University suggest that Centaurs' ability to 'mimic human cognition' may simply be the result of **overfitting**.

In other words, the model may have memorized patterns in the training data instead of actually understanding the task.

To verify this, the research team designed several different experiments. In one experiment, they replaced the original multiple-choice question with a simple instruction: 'Choose answer A'.

If Centaur truly understood the task, the model should have chosen A in every case. But the results show that Centaur still gives the same answers as in the original dataset.

This suggests that the model doesn't truly understand the question, but relies on statistical relationships to arrive at an answer — similar to a student who performs well on an exam because they recognize the question pattern, not because they understand the material.



Lessons for evaluating AI

These findings suggest that evaluating AI models requires greater caution. Large-scale linguistic models are excellent at finding patterns from data, but their 'black-box' nature makes evaluation difficult.

These systems can still encounter familiar problems such as hallucination or misinterpretation of context. Therefore, researchers argue that rigorous and multi-faceted evaluation methods are needed to determine whether AI truly possesses the capabilities it demonstrates.

Despite being described as a cognitive simulation system, Centaur's biggest weakness lies in its ability to understand language. Specifically, the model struggles to grasp the intent of the question.

This shows that achieving true language comprehension—a crucial element in building general cognitive models—remains one of the biggest challenges facing AI today.

The new research doesn't deny the progress of AI, but emphasizes that AI 'thinking like humans' is still quite a long way off.

While AI can mimic human behavior in many situations, that doesn't mean it truly understands. The gap between pattern recognition and genuine thinking remains a crucial line that researchers are trying to bridge.

You finished reading the article "**Accurately assessing the thinking capabilities of AI.**" edited by the [TipsMake](#) team. We hope this article has provided you with many useful tech tips and tricks. You can search for similar articles on tips and guides. Thank you for reading and for following us regularly.