PNAS LETTER

Do large language models have a theory of mind?

Damian K. F. Pang[a,1], Severin K. Y. Pang[b], Marianne D. Broeker[a], and Alix Hibble[a]

Michal Kosinski's recent study on theory of mind (ToM) tasks given to different large language models (LLMs) (1) is fascinating and offers many insights into the continued evolution and development of LLMs.

When testing ToM in animals, much ethological research has focused on differentiating "genuine" ToM from other cognitive functions. *Morgan's Canon* recommends using "lower" rather than "higher" psychological faculties to explain animal behavior where possible (2). While this "canon" may lack justification, "association-blindness" is also problematic (3).

Researchers working in developmental psychology and animal behavior have developed increasingly sophisticated methods to rule out alternative explanations (4), gradually building cumulative cases based on converging evidence (5). We suggest that the same should be done with LLMs. Kosinski considered some alternative explanations and included control trials to exclude simple heuristics (1). We suggest that this should be expanded by examining other alternatives like associative learning, which can be achieved through simple electronic circuits (6) and can be explicitly trained.

Given the sudden jump in performance in newer models, it is likely that LLMs have either been explicitly trained or engineered to solve ToM tasks, which could explain some observed differences from humans (7). Explicit training would likely result in overinferring false beliefs when a similar pattern exists. For example, stating that the container is transparent [inspired by the "goggles experiment" in ethology (8)] should not result in false beliefs while retaining a similar structure to ToM tasks. We suggest that wrongly inferring a false belief in such a scenario would be indicative of explicit training.

LLMs use mathematical representations of word vectors in a multidimensional space that include word associations and positions. Each vector is interpreted through surrounding vectors to give a broader context. Such structured composition can mimic the logic of its training data, given that logical relationships often result in specific vector patterns. LLMs are trained on texts created by humans as well as using reinforcement learning from human feedback (9). Both the training data and the feedback come from humans who possess a ToM, making it at least possible for LLMs to pass ToM tasks simply through pattern recognition.

Testing this would require ToM tasks with radically different patterns (not just novel particulars) from the ones found in the existing literature included in the training data. Alternatively, a significant improvement in ToM task performance in older models through training without model tuning (10) would indicate that patterns in the training data rather than in the model can account for task performance.

None of this implies that LLMs cannot have a genuine ToM. However, we propose that successfully solving isolated ToM tasks is insufficient evidence to indicate the presence of ToM (5). While the studies conducted by Kosinski (1) and others (7) are important and relevant, we suggest that attributing ToM to LLMs may be premature until simpler explanations can be ruled out and a cumulative case based on converging evidence can be made.

Author affiliations: [a]Department of Experimental Psychology, University of Oxford, Oxford OX2 6GG, United Kingdom; and [b]School of Computer Science, University of St. Andrews, St. Andrews KY16 9AJ, United Kingdom

Author contributions: D.K.F.P. convened discussion group; S.K.Y.P., M.D.B., and A.H. participated in discussion group; and D.K.F.P., S.K.Y.P., M.D.B., and A.H. wrote the paper.

The authors declare no competing interest.

[1]To whom correspondence may be addressed. Email: damian.pang@psy.ox.ac.uk.

Published July 3, 2025.

1. M. Kosinski, Evaluating large language models in theory of mind tasks. *Proc. Natl. Acad. Sci. U.S.A.* **121**, e2405460121 (2024).
2. C. L. Morgan, *Comparative Psychology* (Walter Scott, 1903).
3. C. Heyes, Simple minds: A qualified defence of associative learning. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **367**, 2695–2703 (2012).
4. C. Heyes, Animal mindreading: What's the problem? *Psychon. Bull. Rev.* **22**, 313–327 (2014).
5. D. C. Dennett, Beliefs about beliefs [P&W, SR&B]. *Behav. Brain. Sci.* **1**, 568–570 (2019).
6. Y. V. Pershin, M. Di Ventra, Experimental demonstration of associative memory with memristive neural networks. *Neural Netw.* **23**, 881–886 (2010).
7. J. W. A. Strachan *et al.*, Testing theory of mind in large language models and humans. *Nat. Hum. Behav.* **8**, 1285–1295 (2024).
8. K. Karg, M. Schmelz, J. Call, M. Tomasello, The goggles experiment: Can chimpanzees use self-experience to infer what a competitor can see? *Anim. Behav.* **105**, 211–221 (2015).
9. J. Chatterjee, N. Dethlefs, This new conversational AI model can be your friend, philosopher, and guide … and even your worst enemy. *Patterns* **4**, 100676 (2023).
10. X. Liu, T. Pang, C. Fan, "Federated prompting and chain-of-thought reasoning for improving LLMs answering" in *Knowledge Science, Engineering and Management* (Springer Nature, 2023), pp. 3–11.