

UC Merced

Proceedings of the Annual Meeting of the Cognitive Science Society

Title

No Reliable Evidence of Self-Reported Sentience in Small Large Language Models

Permalink

<https://escholarship.org/uc/item/20q7439h>

Journal

Proceedings of the Annual Meeting of the Cognitive Science Society, 48(0)

Authors

Kaiser, Caspar

Enderby, Sean

Publication Date

2026

Copyright Information

This work is made available under the terms of a Creative Commons Attribution License, available at <https://creativecommons.org/licenses/by/4.0/>

Peer reviewed

No Reliable Evidence of Self-Reported Sentience in Small Large Language Models

Caspar Kaiser

Behavioural Science Group, Warwick Business School, University of Warwick

Sean Enderby

Behavioural Science Group, Warwick Business School, University of Warwick

Abstract

Whether language models are sentient cannot currently be determined by direct observation. We therefore ask a more tractable surrogate question: do models report themselves to be sentient, and are these reports truthful? We address this by first querying several open-weights models about their own subjective experience, and then verifying responses using classifiers trained on internal activations. We draw upon three model families (Qwen, Llama, GPT-OSS) ranging from 0.6 billion to 70 billion parameters, approximately 130 questions about subjective experience, and three classification methods from the interpretability literature. First, we find that models consistently deny being sentient: they attribute consciousness to humans but not to themselves. Second, classifiers trained to detect underlying beliefs - rather than mere outputs - provide no evidence that these denials are untruthful. Third, within the Qwen family, larger models deny sentience more confidently than smaller ones. These findings contrast with recent work suggesting that models harbour latent beliefs in their own consciousness.