Autoepistemic Uncertainty in Humans and Large Language Models

Humans routinely express uncertainty by reasoning about what they do and do not know. A standard test of this asks participants to provide numerical confidence intervals that are correct at a specified rate (e.g., 50%). Large language models can generate similar interval judgments, but it is unclear what such responses reflect, given that models may have encountered correct answers during training without access to whether those answers are known or appropriate to report with a given level of confidence. We investigate this issue by comparing human and LLM performance on interval-based uncertainty tasks across questions whose answers vary in prior availability, including well-known facts, obscure but real facts, and plausible synthetic quantities with no ground truth. We use interval estimation to test whether uncertainty judgments reflect epistemic access, as in human reasoning, or instead arise from learned patterns independent of knowledge states, producing apparent autoepistemic behavior through behavioral imitation.

In G. Boleda, I. Dautriche, M. Maldonado, C. Saldana and T. Torrent (Eds.), *Proceedings of the 48th Annual Conference of the Cognitive Science Society*