Bridging Minds and Models: A Comparative Analysis of Human and LLM Reasoning in Think-Aloud Zendo Tasks

An-Chi Chen (anchi.chen.chloe@gmail.com) & Neil Bramley
School of Philosophy, Psychology & Language Sciences, University of Edinburgh

Abstract

Recent large language models (LLMs) can achieve human-like performance on some reasoning tasks, yet it remains unclear whether similar outcomes reflect similar reasoning processes. We compared human think-aloud protocols and prompted GPT-4o chain-of-thought traces on five Zendo reasoning tasks involving relational and compositional rule discovery. Reasoning traces were segmented and coded into predefined cognitive-state categories to examine state usage, transition patterns, and reasoning flow. Although GPT-4o matched human performance on simpler tasks, humans achieved higher overall accuracy (≈60% vs. ≈30%), with larger divergence on tasks requiring more complex relational structure. Process-level analyses showed that human reasoning involved broader state diversity and more frequent cross-state transitions, whereas GPT-4o exhibited more concentrated transition patterns and higher rates of repeated local loops. These findings suggest that process-level analyses can reveal systematic differences in observable reasoning dynamics beyond task accuracy, highlighting the value of cognitive-state and transition analyses for comparing human and model-generated reasoning.

Keywords: human-AI comparison; inductive reasoning; process-level analysis; think-aloud protocols; GPT-4o; chain-of-thought; cognitive-state analysis

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