Human and LLM performance on linguistic test: Content effect and task demands

May Reese
San Francisco State University, San Francisco, California, United States

Anastasia Smirnova
San Francisco State University, San Francisco, California, United States

Abstract

Large Language Models (LLMs) display an impressive set of capabilities in linguistic understanding. While advanced models outperform humans on certain tasks, LLM reasoning and linguistic competency differs from that of humans (Felin & Holweg, 2024; Mahowald et al., 2024; Niu et al., 2024). In this study, we evaluate humans and GPT-4o on the Winograd Schema Challenge, a pronoun resolution task. We focus on Japanese, a relatively understudied language in the emergent field of human-LLM evaluation. To assess human vs. LLM performance, we manipulate task demands and content. We report three findings: (i) Humans outperform LLMs in the baseline condition, i.e. the standard pronoun resolution task. (ii) As task demands increase, both human and LLM performance on the task declines (cf. Hu & Frank, 2024). (iii) We find evidence for content effects (cf. Lampinen et al., 2024): LLMs surpass humans as the content of the task is manipulated to favor LLMs.

In D. Barner, N.R. Bramley, A. Ruggeri and C.M. Walker (Eds.), *Proceedings of the 47th Annual Conference of the Cognitive Science Society*