Human vs. Large Language Model-Based Sampling: Evidence from Large-Scale Replications

Tsz Kwan (Emmy) Wong
Behavioural Science Group, Warwick Business School, University of Warwick, Coventry, United Kingdom

Maximilian Maier
Behavioural Science Group, Warwick Business School, University of Warwick, Coventry, United Kingdom

Abstract

Large language models are increasingly used to model human behaviour, yet debate remains about whether they can meaningfully approximate both average effects and response diversity in human samples. We revisit replication studies from Many Labs 2 and management-science replications using LLM-based samples, comparing direct prompting with silicon sampling. We extend existing silicon sampling by assigning personas using demographic profiles and Big Five traits. Pilot results from ML2 using GPT-4o-mini show that modelling participant heterogeneity increases response variability: silicon-sampling variance was approximately 2.7 times that of standard prompting, exceeded standard-prompting variance in 81% of focal outcomes, and standard prompting produced zero variance in 41% of focal outcomes. In ML2, 52.9% of primary findings were replicated in human data (using significance in the same direction as the replication criterion). Comparing LLM outcomes to ML2 replications, 38.9% of silicon-sampling outcomes showed consistent signals, compared with 22.2% under standard prompting.

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