Recall or Reason? Loss Aversion in Large Language Models

Dr. M. Mahdi Tavalaei

University of Surrey, Surrey Business School, the United Kingdom

Abstract

Large language models (LLMs) are increasingly deployed as decision-makers, advisors, and synthetic participants in social-science research, on the implicit assumption that their underlying statistical architecture lends them a degree of computational rationality that human decision-makers lack. Whether this assumption holds is a foundational question for both behavioural economics and cognitive science: if LLMs inherit the systematic biases that depart from rational choice, the case for treating them as neutral analytical tools weakens substantially. Recent work suggests that they do. Frontier LLMs have been shown to exhibit loss aversion, one of the most robust regularities in human choice, and this finding has been interpreted as evidence that human-like preferences can emerge from large-scale language training. Yet a deeper question remains unresolved: is the apparent bias the product of genuine preference-like reasoning, or a recall of canonical experimental scenarios that are densely represented in LLM training corpora? We test this distinction by eliciting loss-aversion responses from a panel of frontier LLMs across open- and closed-weight families, presenting each model with the canonical scenario alongside structurally identical but novel isomorphs that strip away textbook framing while preserving the underlying choice structure. We further vary the stake size across trivial, canonical, large, and extreme magnitudes to probe regions in which training-data exposure is densest, sparsest, and absent. The findings carry direct implications for the validity of LLM-based behavioural inference.

Keywords: Large Language Models; Loss Aversion; Prospect Theory; Behavioural Economics Of AI; Cognitive Convergence