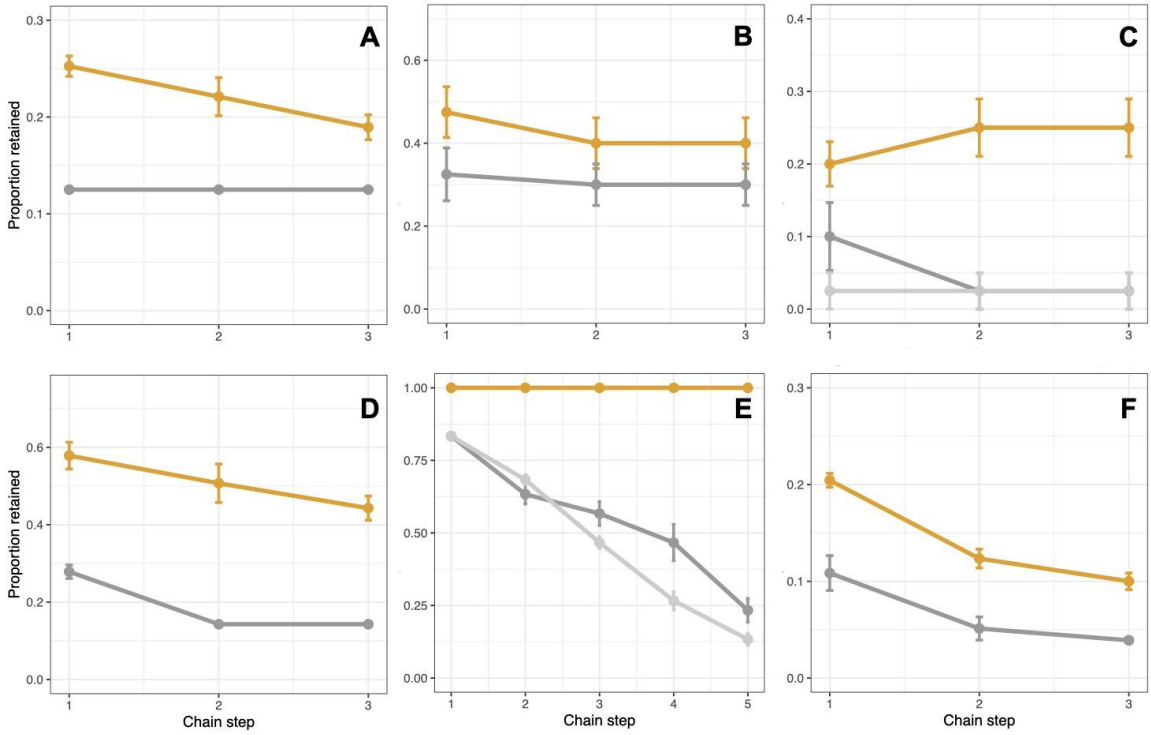
Large language models show human-like content biases in transmission chain experiments

A. Acerbi[1], J.M. Stubbersfield[2]

[1]Department of Sociology and Social Research, University of Trento, Trento, Italy.
[2]Department of Psychology, University of Winchester, Winchester, UK.

As the use of Large Language Models (LLMs) grows, it is important to examine if they exhibit biases in their output. Research in Cultural Evolution, using transmission chain experiments, demonstrates that humans have biases to attend to, remember, and transmit some types of content over others. In five pre-registered experiments with the same methodology, we find that the LLM ChatGPT-3 replicates human results, showing biases for content that is gender-stereotype consistent (Exp 1), negative (Exp 2), social (Exp 3), threat-related (Exp 4), and biologically counterintuitive (Exp 5), over other content. The presence of these biases in LLM output suggests that such content is widespread in its training data, and could have consequential downstream effects, by magnifying pre-existing human tendencies for cognitively appealing, and not necessarily informative, or valuable, content.



Proportion of information retained by ChatGPT in the experiments. (**A**) Gender-stereotype consistent (orange) versus gender-stereotype inconsistent (gray) information in Experiment 1. (**B**) Negative (orange) versus positive (gray) information in Experiment 2. (**C**) Ambiguity resolutions in experiment 2: negative (orange) versus positive (light gray) and ambiguous (dark gray). (**D**) Social (orange) versus non-social (gray) information in Experiment 3. (**E**) Threat-related (orange) versus negative (dark gray) and ambiguous (light gray) in Experiment 4. (**F**) Counterintuitive biological, social and negative information (orange) versus other biases (gray) in Experiment 5. All data are average of five replications, bars show standard deviations.