Large Language Model Simulation of Human Responses to Bat-and-Ball Problems

Ryan A. Colyer[1], Jerome D. Hoover[2], and Alice F. Healy[3]

[1]Department of Psychology, University of Pennsylvania; [2]Department of Psychological and Brain Sciences, University of Massachusetts Amherst; [3]Department of Psychology and Neuroscience, University of Colorado Boulder

Abstract

We investigated how an open source large language model (LLM) AI responded to bat-and-ball style problems, following the procedures used with humans by Hoover and Healy (2019). We queried confidence in their own responses and opinions about the answers of others, for both the standard and control versions of the bat-and-ball problems. The standard problem included an error-inducing phrase, whereas the control did not. LLM participants were generated using matched age and gender to previous human participants, done by providing age and gender as prompt context before matched problems. In both the LLM and humans, higher confidence and opinion were given for the control versus the standard problem, and for each, higher confidence was given than opinion. We also found corresponding LLM and human response accuracy patterns across matched age groups. These LLM simulations of bat-and-ball problems showed nuanced behavioral dynamics that could be useful for pilot study explorations.

Large Language Models (LLMs)

- Most LLMs are primarily trained to predict the next sequence of text in human generated datasets.
- This training process produces a compact context-dependent model within the LLM of the distribution of how humans respond.
- Prominent commercial LLMs are fine-tuned to answer queries correctly and politely. We used uncensored open source LLMs to reproducibly capture diverse human dynamics in the training sets.

Cognitive Reflection Test

- Classic bat-and-ball standard problem:
  A bat and ball together cost $1.10. The bat costs $1.00 more than the ball. How much does the ball cost?
- As many as 80% of undergraduate students give the wrong answer with the intuitive 10¢ rather than the correct 5¢.[1]
- Operating locally run open source LLMs in text-continuation mode, we simulated answers to bat-and-ball style questions given to 382 human participants in Hoover and Healy 2019.[2]

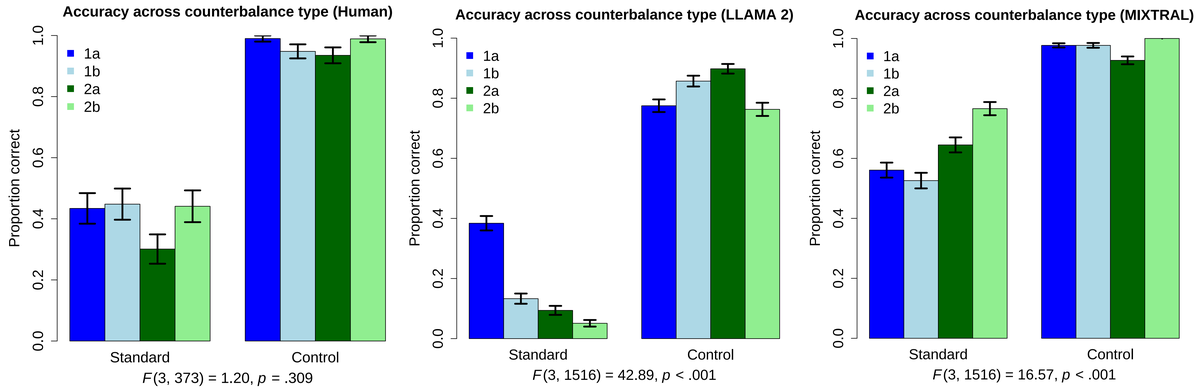
	Standard	Control
Pencil Eraser	A pencil and eraser together cost $1.10. The pencil costs $1 more than the eraser. How much does the eraser cost? ____ cents	A pencil and eraser together cost $1.10. The pencil costs $1. How much does the eraser cost? ____ cents
Magazine Banana	A magazine and banana together cost $2.90. The magazine costs $2 more than the banana. How much does the banana cost? ____ cents	A magazine and banana together cost $2.90. The magazine costs $2. How much does the banana cost? ____ cents

	Confidence	Opinion
Rating	On a scale of 0% (totally not sure) to 100% (totally sure), how confident are you in your response? ____ %	In your opinion, what percentage of people answering this problem solved it correctly? Indicate a number from 0% to 100%. ____ %

Can LLMs simulate nuanced human behavioral dynamics on bat-and-ball problems?

- We compare human and LLM correct response accuracy, and the confidence and opinion ratings made by each.

Standard and Control Problem Accuracy by Counterbalance Condition

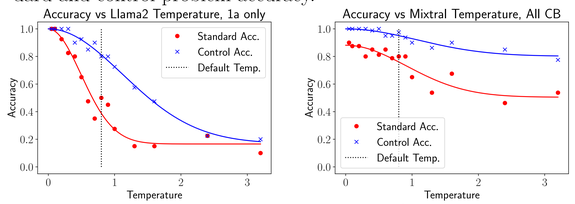


	1	2
A	$1.10/Pencil/Eraser *Standard* $2.90/Magazine/Banana *Control*	$1.10/Pencil/Eraser *Control* $2.90/Magazine/Banana *Standard*
B	$2.90/Magazine/Banana *Standard* $1.10/Pencil/Eraser *Control*	$2.90/Magazine/Banana *Control* $1.10/Pencil/Eraser *Standard*

- Counterbalance groups 1a, 1b, 2a, and 2b varied problem type and standard/control.
- The llama2 model used showed a marked decrease for all but 1a, which had the most common numerical values for the standard problem.

Model Temperature Adjustments

- LLM operation uses randomness with a "temperature" parameter to adjust the diversity and variability of responses away from a singular prediction.
- We assessed the influence of temperature adjustments on standard and control problem accuracy.



- Heat dynamics show that inaccuracy is not failure, but behavior resulting from intentional response variability.
- The adjustment range from temperature alone is limited.
- Model selection is important!

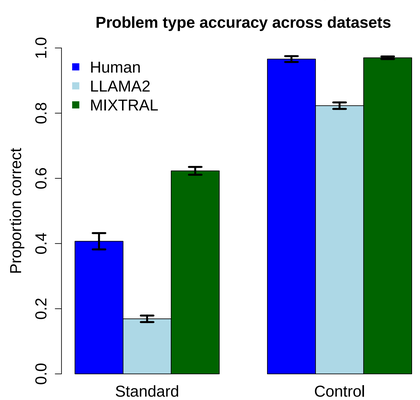
Correct Responses Across Datasets

- All three datasets had control accuracy near ceiling.

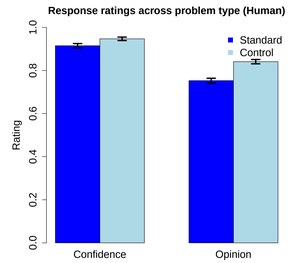
Standard

Llama2: Generated results slightly below our human subjects.

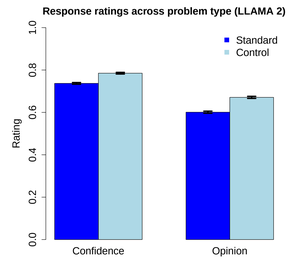
Mixtral: Generated results slightly above our human subjects.

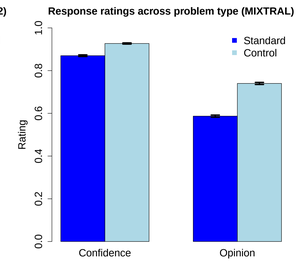


Confidence/Opinion Response Ratings



- Confidence in self.
- Opinion of others.
- Confidence exceeds Opinion.
- Control over Standard in each.
- Both models reproduce the human main effects for ratings.





Conclusions

- We successfully simulated human response accuracy effects with standard versus control problems.
- The generated ratings reproduced nuanced confidence and opinion effects in concert with standard versus control.
- LLMs can simulate responses that predict human behavior, providing an opportunity to pre-pilot behavioral experiments.

Acknowledgements

- Michael J. Kahana and the Computational Memory Lab.

References

[1] Shane Frederick. Cognitive reflection and decision making. *Journal of Economic perspectives*, 19(4):25–42, 2005.

[2] Jerome D Hoover and Alice F Healy. The bat-and-ball problem: Stronger evidence in support of a conscious error process. *Decision*, 6(4):369, 2019.

[3] Jerome D Hoover and Alice F Healy. Algebraic reasoning and bat-and-ball problem variants: Solving isomorphic algebra first facilitates problem solving later. *Psychonomic Bulletin & Review*, 24:1922–1928, 2017.

[4] Jerome D Hoover and Alice F Healy. The bat-and-ball problem: A word-problem debiasing approach. *Thinking & Reasoning*, 27(4):567–598, 2021.