A Warning for AI Biases at Work: Evidence from 40,000 Conversations with Large Language Models

Artificial Intelligence (AI) is increasingly involved in decision-making in work and organizational contexts (Dell'Acqua et al., 2023; Li et al., 2024), and can lead to performance increases (Noy & Zhang, 2023). As AI aims to emulate human thinking (Russell & Norvig, 2010), utilizing it can lead to better work decisions in comparison to human judgment, which often relies on heuristics or is susceptible to biases, stemming from cognitive limitations (Gilovich et al., 2002).

Considering that AI (a) aims to mimic human cognition, which is marked by cognitive biases and heuristics, and (b) has the computational capacity to overcome these biases and heuristics, this research asks, “How do Large Language Models (LLMs) perform when it comes to well-established cognitive biases?” We address this question in six pre-registered experiments testing three judgment heuristics (anchoring, representativeness, and availability; Tversky & Kahneman, 1974) and three biases that violate subjective utility theory (framing effects, endowment effect, and transaction utility; Kahneman et al., 1990).

Across studies, we conducted 40,000 independent trials with GPT-3.5-Turbo-0613 (hereafter, “GPT-3.5”) and GPT-4-0613 (hereafter, “GPT-4”). We developed and used a custom application, written in Node.js, to send requests to, and receive responses from the “Chat Completions” Application Programming Interface (API) of OpenAI (the creator of GPT-3.5 and GPT-4). This API provides a "powered by ChatGPT/OpenAI" service used by many organizations, including most Fortune 500 companies (Porter, 2023). This method allowed us to (a) control for the “temperature” of the models – setting it to “1”, a value that balances determinism and creativity of the responses, (b) sample thousands of responses for both models simultaneously within a few hours reducing potential confounds of time (Chen et al., 2023), and (c) control for the possibility that the GPT model would be trained on the first replies and answer differently later.

Each observation was obtained with a sequence of two prompts. First, we asked the GPT model to assume the role of a research participant. This “virtual participant” was then presented with a study replicting closely a studied heuristic or bias. We repeated the procedure 1,000 times for each condition in each study.

For each study we added a 2×2 experimental design to the core effect under investigation. The first factor of this design was the GPT model version: GPT-3.5 versus the considered as superior GPT-4. The second factor was whether or not the LLM was provided with a 2-sentence “self-debiasing” induction.

We statistically analyzed both the responses given by the LLM and the variance around these responses. This is important because small variances around erroneous responses can increase trust in the results. Across our studies, we find systematic evidence against several assumptions about the performance of LLMs.

First, LLMs did not provide unbiased responses: In all six studies they were biased at least in some conditions, at the .001 level. GPT3.5 was biased in 4/6 studies and GPT-4 in 6/6 studies, but often to smaller degrees.

Second, LLM biases do not systematically replicate human thinking: Either GPT 3.5 (for the endowment effect), or GPT4 (for anchoring and adjustment) produced biases in the reverse direction than humans, at the .001 level. The source of these reverse biases is unlikely to be the training data: If this was the case, they should mimick the direction of human biases.

Third, the more "intelligent" model (GPT4) was biased in 93.2% and 98.7% of the cases for the availability and the representativeness heuristics, respectively. These numbers for GPT3.5 were 19% and 8.5%. The source of these biases is unlikely to be the training method: If this was the case, the better trained model should perform better.

Fourth, LLMs were not unbiased even in cases with an objectively correct answer. For instance, although within its capabilities, GPT4 did not use its training data to see how often the stock market historically went up (our anchoring study question), and it did not scrutinize the provided prompt to find the correct answer for the "Linda" problem (our representiveness heuristic question).

Fifth, the two LLMs systematically differed in dimensions other than their "intelligence".

- In all studies, GPT4 provided much less varied responses (p values < .001). This may be problematic, given that GPT4 was somehow biased in all our studies.
- In the framing study, GPT-3.5 provided overwhelmingly more risk-seeking responses than GPT-4; They respectively chose the risky option 84.93% versus 3.43% of the times.
- In 5/6 studies, with the exception of representativeness, the two AI models handled our 2-sentence debiasing prompt differently: There were significant interactions involving the LLM and the "debiasing" factor at the .001 level.

In conclusion, the results raise a strong warning about the use of AI for work. Across six preregistered experiments and 40000 conversations, the most popular LLMs in work setting gave biased respones, sometimes in unpredictable ways (e.g., opposite to human bias). Our results show that relying on answer consistency or on model "intelligence" to alleviate these biases, may be ineffective. Future research should study how customized LLMs could overcome these limitations.

References

Chen, L., Zaharia, M., & Zou, J. (2023). How is ChatGPT's behavior changing over time? arXiv:2307.09009. https://doi.org/10.48550/arXiv.2307.09009

Dell'Acqua, F., McFowland, E., Mollick, E. R., Lifshitz-Assaf, H., Kellogg, K., Rajendran, S., Krayer, L., Candelon, F., & Lakhani, K. R. (2023). Navigating the Jagged Technological Frontier: Field

Experimental Evidence of the Effects of AI on Knowledge Worker Productivity and Quality. SSRN. https://doi.org/10.2139/ssrn.4573321

Gilovich, T., Griffin, D., & Kahneman, D. (2002). Heuristics and Biases: The Psychology of Intuitive Judgment. Cambridge University Press.

Kahneman, D., Knetsch, J. L., & Thaler, R. H. (1990). Experimental Tests of the Endowment Effect and the Coase Theorem. Journal of Political Economy, 98, 1325–1348.

Li, P., Castelo, N., Katona, Z., & Sarvary, M. (2024). Frontiers: Determining the Validity of Large Language Models for Automated Perceptual Analysis. Marketing Science.

Porter, J. (2023). ChatGPT continues to be one of the fastest-growing services ever. The Verge. https://www.theverge.com/2023/11/6/23948386/chatgpt-active-user-count-openai-developer-conference

Russell, S. J., & Norvig, P. (2010). Artificial Intelligence a Modern Approach. London.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. Science, 185, 1124–1131.