The Future of Large Language Models in Social Science Research: Reply to Berger (2023) and Carrillo, Stachl and Talaifar (2023)

Forthcoming in **American Psychologist**

Sachin Banker

University of Utah

Promothesh Chatterjee

University of Utah

Himanshu Mishra

University of Utah

Arul Mishra

University of Utah

The Future of Large Language Models in Social Science Research: Reply to Berger (2023) and Carrillo, Stachl and Talaifar (2023)

Abstract: In their commentaries, Berger (2023) and Carrillo, Stachl and Talaifar (2023) raise several thoughtful questions regarding machine assisted hypothesis generation in the social sciences. We discuss their ideas and build upon them.

We appreciate Berger (2023) and Carrillo et al. (2023) for their commentary and potential extensions on how generative AI can assist researchers for hypothesis generation. We discuss their ideas and build upon them. Our research provides an initial illustration showing that LLMs are indeed capable of generating novel research hypotheses that can then be tested. We outline directions in which LLMs could accelerate research for the future development of social science.

Berger (2023) raises fundamental questions about whether the current method of human-driven hypothesis generation is in fact ideal. Berger's challenge to traditional, human-centric methods of hypothesis generation prompts a more fundamental reevaluation of the idea-generation process in general. One aspect, as Berger (2023) alludes to, is that traditional human-centric methods might not effectively harness the vast amounts of information available in developing novel research hypotheses. With the ability of LLMs to synthesize information from large volumes of literature that are challenging for humans to keep up with, these newer methods can identify patterns and insights that might be missed by human-centric methods. Advancements in computational techniques have given us tools that can process and analyze information on an unprecedented scale, including harnessing stimuli across multiple domains (such as text, image, audio) and identifying interesting new patterns and questions worthy of further testing from multi-domain data.

Such a democratization of LLMs is expected to progress, driven by reductions in cost associated with the access, fine-tuning, and availability of open-source LLMs. This will aid in offering easy access to prior knowledge. Just as tools like GitHub Copilot have sped up programming tasks—expediting ease and accessibility to code—LLMs can also help democratize the research process. For novice researchers, machine-assisted hypothesis-generating tools can lead to generation of feasible new hypotheses supported by the existing literature; for seasoned researchers, such tools can improve the efficiency with which researchers are able to propose and test new ideas within a field. In line with Sir Isaac Newton’s thoughts, "If I have seen further, it is by standing on the shoulders of giants," LLMs will make it more feasible to stand on the shoulders of giants, see further, and explore new research frontiers.

As Berger (2023) wonders whether such model-generated hypotheses would encourage incremental innovation or produce bigger breakthroughs, it is likely that LLMs like GPT could potentially facilitate both types of hypotheses generation. For incremental innovation, LLMs can efficiently process, analyze, and synthesize vast amounts of literature to refine existing ideas, propose slight modifications, or extend current theories. For more substantial breakthroughs, LLMs can assist researchers by combining knowledge from various disciplines, suggesting novel approaches that might not be readily apparent. Such approaches encourage interdisciplinary work that galvanizes researchers working on, for instance, similar policy-related questions to reach out to their colleagues in diverse domains. But it's crucial to note that the success of these breakthroughs relies on the goals, expertise, and critical evaluation of human researchers.

Carrillo et al. (2023) suggests a more human-centered workflow approach to augment the limitations of both humans and LLMs in the process of hypotheses generation. We agree that the process between researchers and LLMs needs to be collaborative. As researchers continue to actively engage with LLMs, contributing to their refinement and improvement, this engagement will enhance the effectiveness of LLMs, ensuring that they evolve in alignment with the values and practices of the research community. The quality and usefulness of the hypotheses generated by LLMs will depend heavily on the clarity, specificity, and relevance of the prompts provided by the researcher. The process can be improved through the integration of the workflow with scientific databases for automatic cross-verification of generated hypotheses for novelty and similarity with existing work. Moreover, incorporating modules through which the model can learn from researchers' objectives and feedback will contribute to its continuous improvement and mitigation of any preexisting biases. For instance, RLHF (reinforcement learning from human feedback) models are now being used to improve the quality of output (Ouyang et al. 2022). Carillo et al. (2023) offer one example of a workflow that researchers could apply in iteratively evaluating research ideas given certain constraints.

The evidence documented in our research can be considered baseline evidence or a starting point since GPT-4 lacks fine-tuning on specialized domain knowledge. We can expect future advancements with the introduction of LLMs trained on domain-specific knowledge to assist in other research-related tasks with improved accuracy and relevance beyond GPT-4's capabilities. Berger (2023) asks whether generative AI could suggest which model-generated hypotheses are more likely to be valuable or impactful. This is a fascinating but multifaceted issue. Currently generative AI models can identify gaps in the literature by synthesizing existing research and pointing to areas that lack comprehensive study or where contradictions persist. This process can help generate novel hypotheses and compare them with existing theories. However, determining the value or impact of these hypotheses is more complex because it would necessitate evaluating the likely success or broader impact of a hypothesis and require human judgment, expertise in the field, and a nuanced understanding of the research context.

As research landscapes evolve, it's important that methods of hypothesis generation adapt to harness the full potential of available tools, knowledge, and interdisciplinary insights. Our work aims to catalyze a shift in this domain, pushing the boundaries of conventional methods and exploring the potential to accelerate research ingenuity through the synergy of human intuition and machine capabilities.

References

Banker, S., Chatterjee, P., Mishra, H, and Mishra, A. (2023). Machine-Assisted Social Psychology Hypothesis Generation. American Psychologist.

Berger, J. (2023). Machines, Psychology, and Hypothesis Generation: Comment on “Banker, Chatterjee, Mishra, and Mishra (2023). American Psychologist.

Carrillo, A.H., Clemens, S., and Talaifar, S. (2023) A Workflow for Human-Centered Machine-Assisted Hypothesis Generation: Comment on Banker et al. (2023). American Psychologist.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., ... & Lowe, R. (2022). Training language models to follow instructions with human feedback. Advances in Neural Information Processing Systems, 35, 27730-27744.