American Psychologist

ISSN: 0003-066X

2024, Vol. 79, No. 6, 800–802
https://doi.org/10.1037/amp0001256

COMMENTARY

A Workflow for Human-Centered Machine-Assisted Hypothesis Generation: Commentary on Banker et al. (2024)

Alejandro Hermida Carrillo[1], Clemens Stachl[2], and Sanaz Talaifar[3]
[1] Institute for Human Capital Management, Ludwig-Maximilians-Universität München
[2] School of Management, Institute of Behavioral Science and Technology, University of St. Gallen
[3] Department of Management and Entrepreneurship, Imperial College London

Large language models (LLMs) have the potential to revolutionize a key aspect of the scientific process—hypothesis generation. Banker et al. (2024) investigate how GPT-3 and GPT-4 can be used to generate novel hypotheses useful for social psychologists. Although timely, we argue that their approach overlooks the limitations of both humans and LLMs and does not incorporate crucial information on the inquiring researcher's inner world (e.g., values, goals) and outer world (e.g., existing literature) into the hypothesis generation process. Instead, we propose a human-centered workflow (Hope et al., 2023) that recognizes the limitations and capabilities of both the researchers and LLMs. Our workflow features a process of iterative engagement between researchers and GPT-4 that augments—rather than displaces—each researcher's unique role in the hypothesis generation process.

Keywords: large language models, GPT, artificial intelligence, hypothesis generation, human–computer interaction

Artificial intelligence has the potential to revolutionize all aspects of scientific work (Hope et al., 2023), including hypothesis generation. To generate hypotheses, researchers rely on analogic reasoning—the recombination of existing information in novel ways, a capacity that some assert is exhibited by large language models (LLMs) like GPT (Webb et al., 2023). The debate surrounding the exact capabilities of these models is ongoing, and the importance of scrutinizing LLMs' utility in social scientific research generally and psychological hypothesis generation specifically has never been clearer or more timely. To this end, Banker et al. (2024) use GPT-3 and GPT-4 to illustrate how LLMs "can be used as an aid to generate research hypotheses for social psychology" (p. 789). In support of their claim, the authors report that social psychologists rated GPT-generated hypotheses as equal or superior to human-generated hypotheses on dimensions of clarity, originality, impact, plausibility, and relevance. However, in this comment, we recognize the challenges inherent to responsibly and practically incorporating LLMs into the hypothesis generation process. We therefore propose a human-centered workflow that will help researchers leverage the capabilities of LLMs while mitigating the risks that they present (Hope et al., 2023).

Banker et al.'s (2024) approach raises two potential challenges to incorporating LLMs into the hypothesis generation process. First, both LLMs and humans have important limitations that can negatively impact the machine-assisted hypothesis generation process. For example, GPT models tend to present inaccurate or fully made-up information in convincing ways (i.e., to "hallucinate"; Ji et al., 2023). Humans, on the other hand, are prone to confirmation bias (Nuzzo, 2015) and tend to cognitively disengage when interacting with more sophisticated artificial intelligence systems (i.e., "falling asleep at the wheel"; Dell'Acqua, 2021). As a result, using GPT models for

Alejandro Hermida Carrillo https://orcid.org/0000-0002-2882-244X
Clemens Stachl https://orcid.org/0000-0002-4498-3067
Sanaz Talaifar https://orcid.org/0000-0002-4918-9575

The authors have no known conflicts of interest to disclose.

Alejandro Hermida Carrillo wrote the initial draft of this comment. Clemens Stachl and Sanaz Talaifar provided subsequent revisions. Sanaz Talaifar supervised the project.

Correspondence concerning this article should be addressed to Alejandro Hermida Carrillo, Institute for Human Capital Management, Ludwig-Maximilians-Universität München, Geschwister-Scholl-Platz 1, 80539, Munich, Germany. Email: a.hermida@lmu.de

hypothesis generation without appropriate guardrails could lead naïve researchers to waste time and resources pursuing irrelevant research avenues. Second, machine-assisted hypothesis generation that is not informed by researchers' inner worlds (e.g., values, goals) and outer worlds (e.g., discipline-specific controversies) may produce suboptimal hypotheses. For example, Banker et al.'s (2024) approach in Study 2 implies that social psychological scientists have a unanimous desire for "counterintuitive, yet plausible" hypotheses (p. 794). However, some have argued that the pursuit of counterintuitive findings has contributed to low rates of replicability in psychology generally (Hoogeveen et al., 2020) and social psychology specifically (Wilson & Wixted, 2018). It is thus unlikely that all social psychological scientists will find value in such hypotheses. Although Banker et al. (2024) briefly acknowledge that LLMs cannot replace the researcher (p. 796), their demonstration presents LLM-assisted hypothesis generation as a process largely dissociated from the inquiring researcher's goals, values, and context.

We believe that LLMs like GPT-4 can aid in the hypothesis generation process, but only if researchers are aware of and take steps to address the abovementioned issues using best practices for interacting with LLMs (Shieh, 2023; Si et al., 2023). Here, we provide a practical workflow (see Figure 1) for researchers interested in adopting a more human-centered approach to machine-assisted hypothesis generation (Hope et al., 2023). In this workflow, the researcher takes an active role in iteratively providing input to and evaluating output from the LLM.

In the first step of the workflow (1), the researcher initiates a dialogue with GPT-4 using a prompt that is informed by both their desired output and whatever details from their inner worlds (e.g., research interests and goals) and outer worlds (e.g., debates in the field, resource constraints) that they deem relevant for hypothesis generation (see Table 1). In the second step of the workflow (2), GPT-4 responds with tailored and contextualized hypotheses, which the researcher critically evaluates in the third step (3). During this evaluation, the researcher may cross-check GPT-4's suggestions with scientific databases or with their own knowledge. This exercise may reveal that a certain hypothesis has already been explored or is unlikely to be true. The researcher might decide that they simply are not interested in some of the GPT-generated hypotheses or that they would like to have more information on other hypotheses. Informed by this evaluation vis-à-vis their inner and outer worlds, the researcher then reengages with GPT-4 in (1) with a prompt reflecting their new insights. This iterative process of engagement, evaluation, and reengagement continues until the researcher's hypothesis generation goals have been met and they move on to their next task. Throughout the process, the researcher documents and subsequently reports how they used GPT-4 in the hypothesis generation process.

This human-centered workflow for machine-assisted hypothesis generation capitalizes on GPT-4's capabilities for session-based memory (i.e., to maintain conversations) and rapid learning from few examples (Brown et al., 2020). It also recognizes that GPT-4's training data are opaque and time limited, and its output is nonreplicable and unstable (Chen et al., 2023). These features underscore the need for continuous evaluation, verification, and documentation by the researcher. Furthermore, the workflow provides GPT-4 access to each researcher's inner and outer worlds via strategic prompting (see Si et al., 2023), which should produce hypotheses that are more tailored to and thus more valuable to each researcher.

In sum, we illustrate how GPT-4, the state-of-the-art LLM, can be embedded within the scientific pipeline to augment—rather than displace—the researcher's role in hypothesis generation. We see fertile ground for future scholarship demonstrating the complementarity of LLMs and humans in other parts of the research pipeline. In our view, however, such work must take seriously the limitations and capabilities of both the human and the machine.

Figure 1
Workflow for Human-Centered Hypothesis Generation With GPT-4

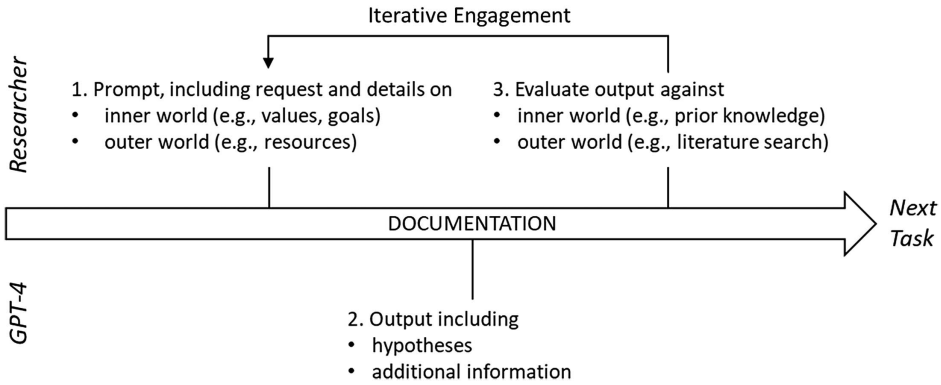

Table 1
Prompt Elements for Hypothesis Generation With GPT-4

Prompt element	Description and examples
Request	
Number and type of hypotheses	___ number (e.g., 5) of _____ (e.g., directional, counterintuitive, testable, plausible, falsifiable) hypotheses
Additional information	Theoretical rationale, explanation of fit with one's research interests/program, keywords to use in a database search
Inner world details	
Topical and/or methodological interests	Verbal descriptions of topical interests (e.g., self and identity, collective action) and methodological interests (e.g., surveys, experiments), potentially with supplementary information provided (e.g., article abstracts, introductions)[a]
Values	Scientific or societal values (e.g., fairness, reproducibility, generalizability, novelty, precision, internal validity)
Current goals	Specific short- or long-term goals concerning a specific research question, study, article, or research program (e.g., developing Study 3 in an article on collective action to preserve cultural heritage)
Outer world details	
Study context	Stage of the project (e.g., early, late), the type of publication, and intended journal/outlet (e.g., empirical article for *JPSP*)
Resources	Assets (e.g., access to samples, data sets, or collaborators) and constraints (e.g., budget)
Knowledge landscape	Recent trends (e.g., collective action is moving online), current debates (e.g., mechanisms linking social identity to collective action)
Discipline-specific issues	Current methodological, conceptual, or theoretical concerns in the field (e.g., importance of non-WEIRD samples)

Note. A nonexhaustive list of elements researchers can include in their prompts to GPT-4 for hypothesis generation, with examples. *JPSP* = *Journal of Personality and Social Psychology*; WEIRD = Western, educated, industrialized, rich, and democratic.
[a] Researchers should be mindful of copyright and privacy issues when inputting data into GPT-4.

References

Banker, S., Chatterjee, P., Mishra, H., & Mishra, A. (2024). Machine-assisted social psychology hypothesis generation. *American Psychologist*, *79*(6), 789–797. https://doi.org/10.1037/amp0001222

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., … Amodei, D. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems*, *33*, 1877–1901. https://doi.org/10.18653/v1/2022.emnlp-main.90

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

Dell'Acqua, F. (2021). *Falling asleep at the wheel: Human/AI collaboration in a field experiment on HR recruiters* [Working paper].

Hoogeveen, S., Sarafoglou, A., & Wagenmakers, E. J. (2020). Laypeople can predict which social-science studies will be replicated successfully. *Advances in Methods and Practices in Psychological Science*, *3*(3), 267–285. https://doi.org/10.1177/2515245920919667

Hope, T., Downey, D., Weld, D. S., Etzioni, O., & Horvitz, E. (2023). A computational inflection for scientific discovery. *Communications of the ACM*, *66*(8), 62–73. https://doi.org/10.1145/3576896

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y. J., Madotto, A., & Fung, P. (2023). Survey of hallucination in natural language generation. *ACM Computing Surveys*, *55*(12), Article 248. https://doi.org/10.1145/3571730

Nuzzo, R. (2015). How scientists fool themselves—And how they can stop. *Nature*, *526*(7572), 182–185. https://doi.org/10.1038/526182a

Shieh, J. (2023). *Best practices for prompt engineering with OpenAI API*. OpenAI. https://help.openai.com/en/articles/6654000-best-practices-for-prompt-engineering-with-openai-api

Si, C., Gan, Z., Yang, Z., Wang, S., Wang, J., Boyd-Graber, J. L., & Wang, L. (2023). *Prompting GPT-3 to be reliable* [Conference session]. The Eleventh International Conference on Learning Representations, ICLR 2023, May 1–5, 2023, Kigali, Rwanda.

Webb, T., Holyoak, K. J., & Lu, H. (2023). Emergent analogical reasoning in large language models. *Nature Human Behaviour*, *7*(9), 1526–1541. https://doi.org/10.1038/s41562-023-01659-w

Wilson, B. M., & Wixted, J. T. (2018). The prior odds of testing a true effect in cognitive and social psychology. *Advances in Methods and Practices in Psychological Science*, *1*(2), 186–197. https://doi.org/10.1177/2515245918767122

Received August 28, 2023
Revision received September 26, 2023
Accepted September 27, 2023 ■