PNAS OPINION

Large Language Models based on historical text could offer informative tools for behavioral science

Michael E. W. Varnum[a,1], Nicolas Baumard[b], Mohammad Atari[c], and Kurt Gray[d]

By training them on various types of historical texts and records, LLMs could help scholars better grasp the mentality of ancient peoples during major historical events. Here, Pericles gives a famous speech around 430 BC, a year after the start of the Peloponnesian War. Image credit: Shutterstock/vkilikov.

The study of human behavior traditionally focuses on the here and now. After all, people cannot take surveys or participate in experiments if they are not alive today. Here, we propose a way to address this limitation—namely, the use of Historical Large Language Models (HLLMs). These generative models, trained on corpora of historical texts, may provide populations of simulated historical participants. In principle, responses from these faux individuals can reflect the psychology of past societies, allowing for a more robust and interdisciplinary science of human nature.

Some have criticized behavioral science for being too parochial, urging expansion beyond participants from so-called WEIRD (Western, Educated, Industrialized, Rich, Democratic) societies (1) (see also https://www.pnas.org/doi/10.1073/pnas.2316690121). Heeding this call, behavioral science now samples more broadly across space, but lacks diversity in time, mostly recruiting from only the present day (2, 3, 4, 5). Developing generalizable theories about human nature requires incorporating data from people beyond those who lived in the past 50–100 years—a small portion of the history of our species. HLLMs trained on corpora produced by past societies offer an escape from this temporal trap, creating new opportunities to gather historical psychological data by simulating the responses of populations that are no longer living. Researchers might, for example, compare the cooperative tendencies of Vikings, ancient Romans, and early modern Japanese in economic games. Or they could explore attitudes about gender roles that were typical among ancient Persians or medieval Europeans. Indeed, HLLMs might provide a novel way to address a wide variety of hard-to-answer questions.

Timely Tool

In broad strokes, Large Language Models (LLMs) are massive neural networks with deep layers, trained on vast amounts of natural language data (i.e., digitized books, social media posts, web pages), that are able to understand and generate natural language output by predicting the most probable word(s) to follow prior words in a sequence (see also The muse in the machine, https://www.pnas.org/doi/full/10.1073/

Author affiliations: [a]Department of Psychology, Arizona State University, Tempe, AZ 85281; [b]Department of Cognitive Sciences, École Normale Supérieure, Paris 75230, France; [c]Department of Psychological and Brain Sciences, University of Massachusetts Amherst, Amherst, MA 01003; and [d]Department of Psychology and Neuroscience, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599

Author contributions: M.E.W.V., N.B., M.A., and K.G. wrote the paper.

The authors declare no competing interest.

Any opinions, findings, conclusions, or recommendations expressed in this work are those of the authors and have not been endorsed by the National Academy of Sciences.

[1]To whom correspondence may be addressed. Email: mvarnum@asu.edu

Published October 9, 2024.

pnas.2306000120). The resulting models are often then fine-tuned through supervised learning to optimize their performance for specific tasks.

Although not originally designed as tools for behavioral science, LLMs (e.g., ChatGPT, Bard, LLaMa) are currently transforming psychology and adjacent fields because they appear to simulate the responses of human participants across many domains. Studies using LLM-simulated participants have successfully replicated human patterns of moral judgment, economic game behavior, cognitive biases, and even the results of Milgram's experiment on obedience (6). These results have led some to propose that LLMs might be used, with some caution, as substitutes for human subjects in behavioral science research (6, 7). A growing number of scientists are using LLM-based agents in psychological paradigms ranging from negotiations, to cognitive tests, to surveys, to behavioral games.

> **LLMs trained on historical texts should enable scientists to more directly measure the responses of diverse past societies by simulating their responses on modern psychological instruments and behavioral measures.**

However, researchers have also acknowledged that LLMs' responses may reflect only a thin slice of human psychological diversity. The data used to train models such as ChatGPT disproportionately sample texts produced by members of WEIRD societies, explaining why the "participant behavior" of ChatGPT in psychological studies may not capture the thoughts and behaviors of people from less-WEIRD cultural groups (8). Thus, current LLMs are inherently limited in their cross-cultural generalizability, but this also implies that LLMs reflect the cultures on whose writings they are trained (9). LLMs trained on different corpora should be able to simulate the psychologies of different cultural groups.

Texts can provide valuable clues to the psychology of the people that create them. Indeed, there is a growing body of research that has used quantitative analyses of word frequencies and co-occurrences in digitized books (e.g., Google Ngrams), newspaper texts, movie dialogue, and other sources to make inferences about psychological tendencies in the past. These datasets have allowed scholars to chart the trajectory of values (10), stereotypes (11), and the strength of social norms (12) within societies based on analyses of natural language used in these cultural products. Indeed, such data sources can be considered "cognitive fossils," reflecting the minds and cultures that produced them (3). Despite the insights gained from these approaches, they are limited because scientists must typically look for fairly indirect proxies of psychological traits and tendencies, rather than measuring them directly. We propose using HLLMs to venture beyond these existing techniques.

Careful Training

LLMs trained on historical texts should enable scientists to more directly measure the responses of diverse past societies by simulating their responses on modern psychological instruments and behavioral measures. Such HLLMs should contain a trace, so to speak, of the collective mentality of the historical people whose writings were used to build them. Just as LLMs trained on text produced by contemporary populations are able to reproduce the psychological responses of those populations, LLMs trained on large corpora of historical text should reasonably reproduce the psychological responses of the historical populations that produced those texts (13). Drawing on data sources including fiction, plays, diaries, letters, and scholarly texts, it may be possible to gain novel insight into the thinking of populations that are no longer living.

HLLMs would enable researchers interested in cultural change to measure the scope and evolution of a wide range of psychological tendencies over a longer time span than just the past century [the era that most current work examines (5)]—studies could potentially focus on trends hundreds of years prior. The ability to simulate the responses of past populations would extend this timeframe, enabling researchers to assess longer-term dynamics of cultural change.

For example, one could estimate levels of individualism or intergroup prejudice across time by administering psychological scales to samples of simulated participants from adjacent historical eras. And one could generate these estimates of psychological tendencies somewhat more directly than approaches that use word counting or natural language processing of historical texts to make such inferences—namely, through responses on surveys or psychological scales with simulated participant samples from past eras.

HLLMs might also be used to test the historical generalizability of psychological phenomena observed across contemporary societies, including the importance of kin care vs. other fundamental social motives (14), the tendency of people to cooperate more with in- vs. out-group members (15), and sex differences in mating strategies and preferences (16). If replicated with HLLMs, this would add weight to current claims regarding the evolutionary origins of these apparently universal tendencies. If not, then this might suggest a problem in theory and a need to modify the prevailing explanations for these effects.

HLLMs provide opportunities to promote new connections with other fields, especially history and cultural evolution. So far, historians have often been limited to relatively indirect proxies, such as private letters, individual portraits, philosophical treatises, or epic poetry, frequently analyzed using qualitative approaches (17, 18). Historical LLMs could complement such efforts by enabling scholars to simulate and generate somewhat more direct quantitative data on traits, attitudes, and behavioral tendencies.

Researchers have already begun to use LLMs to attempt to infer the psychology of people living in the past. One early example is MonadGPT (for code and description, see ref. 19), a fine-tuned LLM trained on 11,000 early modern historical texts in English, French, and Latin, which relies on historical references in conversation mode. Early demonstrations suggest that MonadGPT may indeed reflect the mentality of early modern Europeans. For example, this LLM responds to questions about the solar system in a manner consistent with the knowledge of Europeans in the 17th century and provides medical advice based on the four-humors model predominant at the time. Another recent LLM is XunziALLM (for code and description, see ref. 20), which was trained on ancient Chinese texts and can generate poetry that matches the rules and forms typical of ancient Chinese poetry. And researchers have recently used language models such as SBERT and GPT on the Chinese Historical Psychology Corpus (C-HI-PSY) to study the values of 11th-century Chinese officials as expressed in their writings; they found a high degree of agreement between expert-rated levels of attitudes toward reforms and the computationally derived estimates of traditionalism (13). These models are still in an experimental phase, and, to some degree, they're built on top of existing LLMs developed with largely contemporary text sources. Nevertheless, it's possible that the models could be used to simulate historical samples for research.

Although these are encouraging first steps, it's still unclear how much these approaches accurately simulate the psychology of past populations. For example, the responses of a modern LLM, even one fine-tuned on historical texts, may reflect modern psychological tendencies or contemporary stereotypes about past populations, rather than the true underlying mindset of those groups.

How might we build a truly historical LLM? To start with, one would need to acquire a sizable amount of historical text from a society from a specific time period. Next, the text needs to be converted to a machine-readable format. Then, it is turned into encoded vectors, fed to a neural network architecture, used to generate probability distributions for words, and, finally, an LLM. Once such a model has been created, using a chat interface, one can simulate participants and run them through psychological experiments (Fig. 1).

Challenges and Caveats

Although the concept of HLLMs is fairly straightforward, and although they have a great deal of promise for enriching the social sciences, there are several challenges for those who would create and deploy them.

Users will have to acquire sufficient training data. Unlike current LLMs, HLLMs cannot draw on billions of internet pages, social media posts, and digitized books. Training corpora will likely be significantly smaller, possibly reducing the quality of the responses that such models generate. However, historical corpora are becoming larger and increasingly available. For example, a French team created an open-access corpus of historical French language texts containing 85 billion words. In the future, we anticipate that the amount of text available to train HLLMs is likely to increase substantially and likely come from a greater variety of world languages.

And further computational advances may enable the creation of LLMs that require smaller training sets to achieve the type of performance that current LLMs are capable of (21). LLMs such as ChatGPT and LaMDA were built from terabytes of data, and in the case of GPT, hundreds of billions of words. However, recent advances show that LLMs can be constructed from datasets that are far smaller—approximately 7 billion words in the case of Microsoft's phi-1 LLM (22). With the training dataset size less of a constraint, creating LLMs from historical corpora should be technically feasible relatively soon.

Another challenge is benchmarking (2). When conducting studies with simulated participants based on current LLMs, it's possible to validate one's results with data from contemporary human participants. Given that data from psychological experiments and surveys generally go back no further than a century or so, these benchmarks will not be available for HLLMs aiming to simulate the responses of more ancient populations. One possibility may be to use historical psychology to quantify tendencies such as romantic love, parental love, and interpersonal trust (3), as well as views on political reform (13), over different historical time periods. Further, ethnographic data recorded over the last few centuries can also be used to benchmark the outputs of HLLMs against a ground truth (23).

Researchers might also query an HLLM regarding its understanding of cosmology, biology, or factual events to assess whether that understanding matches the knowledge that a given population possessed in a past era. Experts on a historical population could use these various approaches to assess the historical accuracy of an HLLM and to fine tune it. Thus, although benchmarking an HLLM may present some unique challenges, there are a number of possibilities for doing so

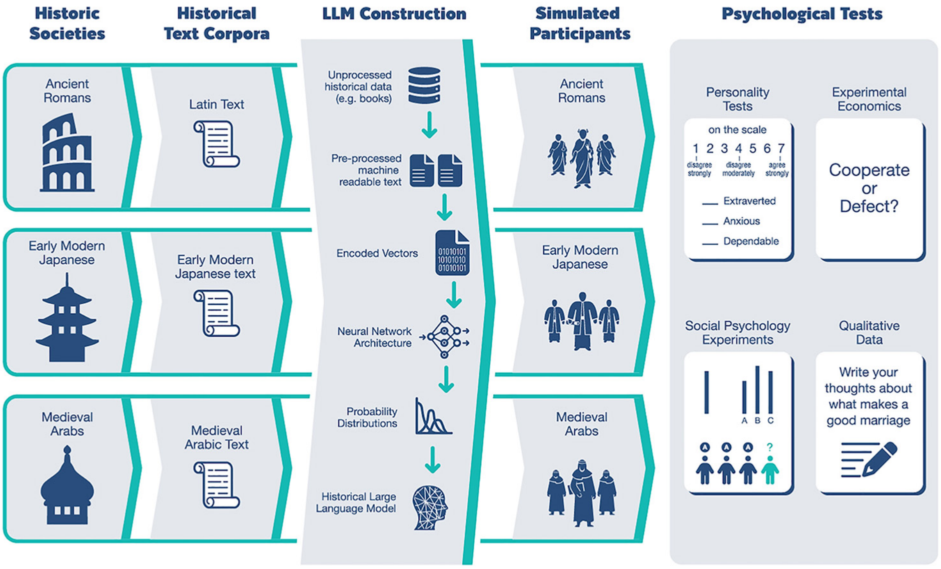

Fig. 1. A conceptual model for the construction and use of HLLMs.

that can increase our confidence that these models are reflecting the psychology of the past population in question.

All LLMs are a product of their training corpora, and HLLMs face challenges in terms of sampling, given that surviving historical texts are likely not representative samples of people who lived in a particular period. Before the advent of widespread literacy in the modern era, a much smaller, and less representative, proportion of the population—the elite—was able to produce written text that could be used to train these models. As a result, it could be hard to generalize from these models.

Any analysis of historical text that intends to draw conclusions about the mentality of past populations faces similar limitations. Educated elites from the societies in question are overrepresented, and their psychology may differ in important ways from that of nonelites (24–26). To help address this limitation, researchers using HLLM approaches can validate their results by drawing from other archival sources, including economic data, public records of marriage and divorce, and data on patents and legislation much in the same way that studies using Google Ngrams have done (10, 12, 27).

In a similar vein, benchmarking using more traditional approaches, including qualitative historical analyses, ethnography, and archaeological data, will be an important step in validating research using HLLMs. And it will help to ensure that the traces of past psychology they capture do indeed represent more than just the thinking of elites. Perhaps HLLMs could also be fine-tuned by weighting some inputs, or making statistical adjustments to HLLMs' aggregate responses on surveys or tests, according to what we know about effects of socio-economic status among modern populations (24–26). Indeed, HLLMs will likely be most effective when used in combination with other approaches.

We acknowledge that our proposal is forward-looking and that creating HLLMs will come with substantial challenges. But we believe that such models will ultimately not only complement current approaches to inferring the mentality of past populations, but also may provide flexible novel research tools that yield important insights beyond what is possible with either traditional empirical social science methods or qualitative approaches in history and philology. Ultimately, we hope that such tools may help us to gain novel insight into the psychology of an understudied pool of humans—namely, the dead (2).

ACKNOWLEDGMENTS. The authors thank Julie Bertoni for her insightful suggestions for dealing with HLLMs potential sampling bias.

1. J. Henrich, S.J. Heine, A. Norenzayan, The weirdest people in the world? *Behav. Brain Sci.* **33**, 61–83 (2010).
2. M. Atari, J. Henrich, Historical psychology. *Curr. Directions Pyschol. Sci.* **32**, 176–183 (2023).
3. N. Baumard, L. Safra, M. Martins, C. Chevallier, Cognitive fossils: Using cultural artifacts to reconstruct psychological changes throughout history. *Trends Cogn. Sci.* **28**, 172–186 (2024).
4. M. Muthukrishna, J. Henrich, E. Slingerland, Psychology as a historical science. *Annu. Rev. Psychol.* **72**, 717–749 (2021).
5. M. E. W. Varnum, I. Grossmann, Cultural change: The how and the why. *Perspect. Psychol. Sci.* **12**, 956–972 (2017).
6. D. Dillion, N. Tandon, Y. Gu, K. Gray, Can AI language models replace human participants? *Trends Cogn. Sci.* **7**, 597–600 (2023).
7. I. Grossmann *et al.*, AI and the transformation of social science research. *Science* **380**, 1108–1109 (2023).
8. M. Atari, M. J. Xue, P. S. Park, D. Blasi, J. Henrich, *Which humans?* PsyArXiv [Preprint] (2024). https://doi.org/10.31234/osf.io/5b26t (Accessed 21 July 2024).
9. N. Buttrick, Studying large language models as compression algorithms for human culture. *Trends Cogn. Sci.* **28**, 187–189 (2024).
10. I. Grossmann, M. E. W. Varnum, Social structure, infectious diseases, disasters, secularism, and cultural change in America. *Psychol. Sci.* **26**, 311–324 (2015).
11. N. Garg, L. Schiebinger, D. Jurafsky, J. Zou, Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proc. Natl. Acad. Sci. U.S.A.* **115**, E3635–E3644 (2018).
12. J. C. Jackson, M. Gelfand, S. De, A. Fox, The loosening of American culture over 200 years is associated with a creativity–order trade-off. *Nat. Hum. Behav.* **3**, 244–250 (2019).
13. Y. Chen, S. Li, Y. Li, M. Atari, Surveying the dead minds: Historical-psychological text analysis with contextualized construct representation (CCR) for classical Chinese. arXiv [Preprint] (2024). https://arxiv.org/abs/2403.00509 (Accessed 7 September 2024).
14. C. M. Pick *et al.*, Family still matters: Human social motivation across 42 countries during a global pandemic. *Evol. Hum. Behav.* **43**, 527–535 (2022).
15. A. Romano, M. Sutter, J. H. Liu, T. Yamagishi, D. Balliet, National parochialism is ubiquitous across 42 nations around the world. *Nat. Commun.* **12**, 4456 (2021).
16. D. P. Schmitt, Sociosexuality from Argentina to Zimbabwe: A 48-nation study of sex, culture, and strategies of human mating. *Behav. Brain Sci.* **28**, 247–275 (2005).
17. P. Ariès, G. G. Duby, Eds., *A History of Private Life: From Pagan Rome to Byzantium* (Harvard University Press, 1987), **vol. 1**.
18. P. Burke, *What is Cultural History?* (John Wiley & Sons, 2019).
19. HackerNews, "MonadGPT - What would have happened if ChatGPT was invented in the 17th century?" (2023). https://news.ycombinator.com/item?id=38407510
20. XunziALLM /README_en.md. GitHub. https://github.com/Xunzi-LLM-of-Chinese-classics/XunziALLM/blob/main/README_en.md
21. R. Eldan, Y. Li, *TinyStories: How small can language models be and still speak coherent English?* arXiv [Preprint] (2023). https://arxiv.org/abs/2305.07759 (Accessed 16 April 2024).
22. S. Gunasekar *et al.*, Textbooks are all you need. arXiv [Preprint] (2023). https://arxiv.org/abs/2306.11644 (Accessed 21 July 2024).
23. D. Bahrami-Rad, A. Becker, J. Henrich, Tabulated nonsense? Testing the validity of the Ethnographic Atlas. *Econ. Lett.* **204**, 109880 (2021).
24. M. W. Kraus, P. K. Piff, R. Mendoza-Dutton, M. L. Rheinschmidt, D. Keltner, Social class, solipsism, and contextualism: How the rich are different from the poor. *Psychol. Rev.* **119**, 546–572 (2012).
25. M. E. Varnum, S. Kitayama, The neuroscience of social class. *Curr. Opin. Psychol.* **18**, 147–151 (2017).
26. G. V. Pepper, D. Nettle, The behavioural constellation of deprivation: Causes and consequences. *Behav. Brain Sci.* **40**, e314. (2017).
27. M. E. W. Varnum, I. Grossmann, Pathogen prevalence is associated with cultural changes in gender equality. *Nat. Hum. Behav.* **1**, 0003 (2017).