Studying Large Language Models as Compression Algorithms for Human Culture

Nicholas Buttrick

University of Wisconsin-Madison

Author Note:

Nicholas Buttrick, Department of Psychology, University of Wisconsin-Madison; Correspondence concerning this article should be directed to Nicholas Buttrick, Department of Psychology, University of Wisconsin-Madison, Brogden Psychology Building, 1202 W Johnson St, Madison, WI 53706. E-mail: nbuttrick@wisc.edu

Abstract:

Large language models extract and reproduce the statistical regularities in their training data. Researchers can use these models to study the conceptual relationships encoded in this training data (i.e., the open internet), providing a remarkable opportunity to understand the cultural distinctions embedded within much of recorded human communication.

Keywords:

Cultural psychology; large language models; compression

A large language model (LLM) is a neural network that uses immense computing resources and training data (along with human-annotated feedback for fine-tuning) to compellingly reply to natural-language prompts. Research exploring the ways that LLMs can contribute to and benefit from psychological understanding has become a central topic in the behavioral sciences [1], and researchers have demonstrated that LLMs, without additional training, can mimic human-like attitudes, reasoning strategies, and biases [2].

Beyond debates about whether LLMs truly 'understand' language or the conceptual spaces they seem to utilize in their responses [3], the massive amount of training data underlying contemporary LLMs provides a remarkable opportunity: studying the output of large-language models themselves as a way of understanding differences in human cultures. The writer Ted Chiang has memorably described ChatGPT as a 'blurry JPEG of the web'[i]: In identifying statistical regularities in its training data and then reproducing them when queried, LLMs act as a sort of compression algorithm for massive corpora of texts. Like with other lossy compression algorithms, an LLM tries to extract the key information from its underlying data and preserve the relationships within texts, even if it cannot perfectly regenerate all its training data. By carefully querying an LLM trained on a large portion of the open web, we can use the ensuing compressed artifact to understand the ways that concepts are statistically co-related in a large portion of all the text ever written.

Searchable compressed artifacts have proven to be invaluable to scholars of culture by facilitating the robust study of differences in cultural production. In the past decade or so, for example, the study of the Google Books nGram corpus, which simply lets researchers track the frequency of words in a large number of texts from 1800 to the present day, has provided cultural psychology with a useful toolkit (see [4] for a review). A basic compression of the majority of books written in English over the past two centuries has been able to meaningfully extend the study of culture. How much more powerful would it be to be able to easily query the concepts embedded in the modern internet, a corpus far larger and more expansive?

LLMs Can Reproduce Cultural Differences

To be an effective tool for studying culture, an LLM as a compressed artifact needs to be able to store and display different cultural worldviews. Cross-cultural analysis of GPT-4 suggests that, when queried by default, it reproduces the mainly White, European, Educated, Industrialized, and Democratic (i.e. WEIRD) psychology that makes up the majority of its training data [5]. There is some evidence, however, that LLMs can indeed capture real-world group differences. Researchers have started to explore the use of personal prompting - asking LLMs to take on certain roles by assigning the LLM to respond 'as if' they were a person with certain properties, such as age, expertise, or personality [6]. When prompted to take on political and demographic roles, LLMs have been surprisingly accurate in their recreation of ground-truth attitudinal differences between American Democrats and Republicans [7] and between various American demographic groups [8].

In our own lab, we have used LLMs to explore one aspect of cultural psychology: the difference between living in urban or rural areas. By prompting LLMs to answer questions core to the psychological study of culture such as 'Who am I?,' 'What is an ideal life?,' or 'What are the qualities of a good friend?,' taking on the role of either someone in a big city or in a small town, we too find that the responses provided by a demographically-prompted LLM line up with what psychological theory (e.g., 9) would predict, finding, for example, that an LLM asked to take on the role of someone from an urban area is much more likely to mention psychological richness as part of a good life than an LLM prompted to take the perspective of someone from a rural area. Through very simple prompting, we can recreate meaningful cultural differences in how groups understand their worlds.

LLMs Are Biased Stenographers

Researchers do need to be circumspect in using this approach. For one, of course, expression on the internet does not capture all of a given culture, no matter how faithfully-transcribed. Much of the way that

a culture comes together is non-linguistic, non-overt, or otherwise opaque to the analysis of any linguistic corpus, and an LLM can only make use of what elements of culture make it into its training data. The things that people choose to put on the internet, moreover, are not faithful transcriptions of the world; rather they are shaped by a myriad of forces, such as literacy, power, and the rewards that come with posting, and these forces must be considered when interpreting the outputs of LLM models. When it comes to the analysis of compressed artifacts, much has been written about problems with the uncritical use of the Google nGram artifact and how failing to think clearly about the underlying data sources (such as the way that the corpus weighs obscure books equally with blockbusters; or the marked rise in scientific texts in the corpus) can lead researchers towards biased conclusions about the world (e.g., [4, 10]).

Similar issues are at play with the use of LLMs for culture: there are major limitations in the data that LLMs are trained on, and therefore what parts of the internet they are actually compressing. Biases in the training data will bias the ability of an LLM to accurately represent group-level attitudes and beliefs: researchers have documented that the current filtering processes used to train LLMs seem to prioritize linguistic content that is often a marker of higher social class [11]; LLMs can represent English-language content far better than non-English content [12]; and LLMs, learning from the biased training data that is the contemporary internet, may be more likely to negatively stereotype or exoticize minorities in their prompted responses [13]. Perhaps unsurprisingly then, prompted LLMs are better at representing ground-truth attitudes of American demographics that are better represented on the internet than groups, such as older Americans, whose internet use is less voluminous [8]; and may be better at representing WEIRD psychology than the psychology of the rest of the world [6].

Further care should be taken to unpack the differences between writing and being written about. Cultural groups can be represented both through their own communications as well as through the writings of outside observers, who may be prone to stereotyping or may have other biases in the way that they

describe attitudes, beliefs, and practices of particular outgroups (see [14] for a further discussion). As LLMs ingest material on the internet without reference to the speaker, it may be that they overrepresent stereotypes of given groups, especially where outgroup writings are more voluminous than writings from inside the group.

The training data may bias the interpretation of LLMs in another way: it may be the case that LLMs are learning from the cultural psychology literature itself and then repeating back to us the things that psychologists have already discovered - that, in other words, the extant scientific literature is highly influential in the model output (e.g., [15]).

An LLM-based approach to investigating cultural representation is especially powerful where it is hard to otherwise survey the group in question or where there is little prior academic data to go on, but these are also the places where it can be hardest to know if the LLM is faithfully representing the opinions and beliefs of a set of people or if, instead, the LLM is representing stereotypes of the group on the internet. Understanding how given groups are represented in the broader culture is still interesting of course, but it is not necessarily the same as understanding ground-truth differences in how groups understand the world themselves. LLMs are likely best used as a spur for further research, therefore, not as the last word on understanding cultural similarities and differences. See Box 1 for additional open questions.

Studying the Internet via LLM is an Opportunity

The study of LLMs provides an opportunity to improve our understanding of cultural representation. By allowing researchers a tractable front-end querying system for the concepts encoded in the internet, they potentially allow for the study of culture at a massive scale. The internet is a tremendously important feature of contemporary culture - it is increasingly where we see each other, present ourselves, and learn about the world. The study of the compressed-internet-artifact embedded within large language models offers researchers a systemic peek into that world. If used carefully and judiciously, the study of LLMs can help us

investigate how cultures are represented within the modern internet, the largest corpus of ideas ever collected.

Box 1: Open Questions

At what level of granularity are LLMs best suited for capturing cultural differences? LLMs are only as good as their training data, and it is not a priori clear how to assess how well a group needs to be represented in the internet for an LLM to pick up on its cultural distinctiveness. What is the smallest, most precise sort of group that an LLM can reliably channel, and what factors affect how trustworthy that channeling can be? Do the models accurately reproduce intersectional identities? Does the level of granularity differ as cultures depart from WEIRDness?

Can LLMs reliably reproduce cultural distinction in non-linguistic modalities? Culture is often encoded in visual or aural media - do LLMs have the power to decode cultural differences in these a-lingustic formats, or are they limited only to reproducing what can be reducible to language?

Does the specific LLM matter? Different LLMs have different underlying training sets and different approaches to fine-tuning. Do these different models, such as LLaMA, GPT, Bard, BLOOM, or Ernie have differential strengths or weaknesses in identifying or reproducing cultural distinctiveness?

Resources:

[i] Chiang, T. (2023, February 9). ChatGPT is a blurry JPEG of the Web. The New Yorker. https://www.newyorker.com/tech/annals-of-technology/chatgpt-is-a-blurry-jpeg-of-the-web

References:

1. Demszky, D. et al (2023). Using large language models in psychology. *Nature Reviews Psychology, 2*, 688–701. https://doi.org/10.1038/s44159-023-00241-5.

2. Dillion, D. et al. (2023). Can AI language models replace human participants?. *Trends in Cognitive Sciences*. https://doi.org/10.1016/j.tics.2023.04.008

3. Mitchell, M., & Krakauer, D. C. (2023). The debate over understanding in AI's large language models. *Proceedings of the National Academy of Sciences, 120*(13), e2215907120.

4. Younes, N. & Reips, U. D. (2019). Guideline for improving the reliability of Google Ngram studies: Evidence from religious terms. *PloS one, 14*(3), e0213554.

5. Atari, M. et al (2023). Which humans? *PsyArxiv.* Published online September 22, 2023. https://doi.org/10.31234/osf.io/5b26t.

6. Kovač, G. et al. (2023). Large language models as superpositions of cultural perspectives. *arXiv.* Published online November 7, 2023. https://doi.org/10.48550/arXiv.2307.07870.

7. Argyle, L. P. et al. (2023). Out of one, many: Using language models to simulate human samples. *Political Analysis, 31*(3), 337-351.

8. Santurkar, S. et al. (2023). Whose opinions do language models reflect?. *arXiv.* Published online March 30, 2023. https://doi.org/10.48550/arXiv.2303.17548

9. Thomson, R. et al. (2018). Relational mobility predicts social behaviors in 39 countries and is tied to historical farming and threat. *Proceedings of the National Academy of Sciences, 115*(29), 7521-7526.

10. Gooding P. (2012). Mass digitization and the garbage dump: The conflicting needs of quantitative and qualitative methods. *Literary and Linguistic Computing, 28*(3), 425–431.

11. Gururangan, S. et al. (2022). Whose language counts as high quality? Measuring language ideologies in text data selection. *arXiv.* Published online January 25, 2022. https://doi.org/10.48550/arXiv.2201.10474.

12. Schott, T. et al. (2023,). Polyglot or not? Measuring multilingual encyclopedic knowledge retrieval from foundation language models. *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing,* 11238–11253. https://doi.org/10.18653/v1/2023.emnlp-main.691.

13. Cheng, M. et al. (2023). Marked personas: Using natural language prompts to measure stereotypes in language models. *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, 1504-1532. https://doi.org/10.18653/v1/2023.acl-long.84

14. Said, O. (1973). *Orientalism*. Vintage.

15. Grosse, R. et al. (2023). Studying large language model generalization with influence functions. *arXiv.* Published online August 7, 2023. https://doi.org/10.48550/arXiv.2308.03296.