Digital Society (2025) 4:51
https://doi.org/10.1007/s44206-025-00209-3

BRIEF COMMUNICATION

Generative AI - Signal to Noise

Adam Kane[1] · **Ricardo Correia**[2] · **Kevin Healy**[3] · **Andrew Jackson**[4]

Received: 4 April 2025 / Accepted: 3 June 2025 / Published online: 20 June 2025

Abstract
The sudden deployment of large language models (LLMs) has been a seismic event for science, with professional scientists, including biologists, struggling to work out how to fit this new technology into their working lives. The benefits of LLMs are manifold but here we flag a neglected and very serious negative aspect of their use in the area of culturomics. This field depends on analysing word frequencies to pick out the prevailing zeitgeist in corpora of text that are readily available online through social media and analysable through modern software. This provides insights into human culture on a scale that was impossible 20 years ago. Culturomics has influenced many topics where understanding the human perspective is key. However, LLMs are 'polluting the waters' by producing AI generated text that is, by definition, not what people are talking about. We believe there's a strong case to be made for highlighting the nature of LLM pollution and give our view for how to clean the waters.

Keywords AI · Culturomics · Media · LLMs · Text analysis

1 Main Body

The combination of digitised text, computational tools and the Web has provided a wealth of readily accessible data for 21st century researchers of human culture to draw on. People in fields as diverse as political science and conservation biology

✉ Adam Kane
adam.kane@ucd.ie

[1] School of Biology and Environmental Science, University College Dublin, Dublin, Ireland

[2] Biodiversity Unit, University of Turku, Turku, Finland

[3] School of Natural Sciences, University of Galway, Galway, Ireland

[4] Department of Zoology, Trinity College Dublin, Dublin, Ireland

have teased out trends from what people are saying and writing about by analysing changes in word frequencies across vast corpora of text. Moreover, the explosion of social media has meant that billions of viewpoints can be mined for insights. The success of this approach has prompted the development of a new field – culturomics – comparable to genomics or proteomics but with a focus on units of culture, typically text, rather than genes or proteins (Ladle et al., 2016). Culturomics can reveal everything from how recommendation algorithms reinforce the culture wars (Salles et al. 2023), how public sentiment is impacted by economic policy (Karami, Bennett, and He 2018) and how netizens respond to the media portrayal of species (Fernández-Bellon & Kane, 2020). Such insights would have been impossible to garner only a couple of decades ago as the time taken to analyse text on a comparable scale would have been far too time intensive.

Culturomics is not without its challenges. Chief among them are the biases inherent in many of the corpora of text that make up their foundation. Google N-grams include an extensive but nevertheless incomplete sample of text and cannot capture the popularity of the work it stores, and hence inherent biases already exist in the sample. A prolific unread author will have more of an influence than a successful one-shot wonder (Zhang, 2015). Wikipedia tends to be left leaning in its articles when it comes to politics (Rozado, 2024). Then there are whole cultures and languages left out of social media analysis if a given country does not have access to a platform or it's simply unpopular among the residents. Indeed, an English language bias has been identified multiple times in the conservation literature (Amano and Berdejo-Espinola 2024). But with all of this said, the majority of this text is still the product of a person, and these biases can be corrected for.

More serious again is the proliferation of automated spam bots whose presence and spread of text can spoil the underlying data in a way that would undermine any inference about human culture. There is something of an arms race between detectors and spammers (Cresci et al., 2017) but, once again, there have been means to identify social media bots.

Enter generative AI. The fanfare surrounding ChatGPT and its ilk shows no sign of abating, and this should be cause for concern among the culturomics community. Large language models (LLMs) function by drawing on digitized text and 'learning' the relationships among our words. In so doing LLMs can generate text that is now generally indistinguishable from human-generated language. The result has been an explosion in artificial text that is representative of the training set the LLMs were based on. As one commentator lamented "Now the Web at large is full of slop generated by large language models, written by no one.

to communicate nothing. Including this slop in the data skews the word frequencies." (McQuater, 2024).

Historically, science has faced and continues to face a diversity of cases where data have been either literally or figuratively polluted. Consider, for example, the impact of nuclear detonations on isotope analysis or the long history of scientific forgeries and fabrications that exact an opportunity cost on researchers to uncover, highlight, and remove from the record.

The worry with LLMs is that there's no immediately obvious solution to cleaning the Web of their "slop" or even identifying it, which makes it qualitatively different

from the above-mentioned pollutants. LLM output isn't watermarked, it's effectively indistinguishable from human text in most cases because it's learned from how we communicate, and the various models are freely available to anyone interested in signing up for an account. A broad-brush worst-case scenario is one in which artificial text outpaces human language with no way to discern between the two and so much the worse for culturomics and related fields. What we'd be left with is a case of Ouroboros, the serpent who ate its own tail. The LLMs won't be able to learn from humans because the model training data is generated by the models themselves. Some concerning impacts are already being felt e.g. with the wordfreq database no longer being updated[1].

We can discern more nuanced implications depending on whether LLMs shape the *style* of communication, as has been observed in academic papers where certain style words have increased in step with AI development, or if broader *topics* are affected. The former case may impact linguistics research mostly, but one would probably still pick up largely similar trends of keywords associated with the main topics of interest (e.g. biodiversity or climate change). This phenomenon where style is the affected aspect shows that we currently have a means to determine to some extent AI-derived content (Kobak et al. 2024), at least for now. However, the latter would have a much broader impact on a wider number of fields that draw from culturomics. Indeed, this could effect a cultural homogenization, a fear that echoes Jaron Lanier's earlier concerns about a "digital Maoism" when Wikipedia was establishing itself as the acme of online information (Lanier, 2006).

A more dramatic result could arise if some bad actor crafted a LLM with their nefarious aims in mind. Imagine a LLM trained in a way that specifically ignores or amplifies topics of controversy or nuance which is then let loose to populate the web with its output e.g. Grok's recent singular focus on 'white genocide'. Digital Maoism indeed.

The immediate response in our view should be to highlight which corpora (i.e., collections of written text) are likely most prone to being 'contaminated' by LLM output and which are relatively safe. This would give researchers *some* confidence that their inference is based on the thoughts of people with editorial oversight being a key feature. Here we offer a few examples along the spectrum. The output of professional journalists is likely to have strong editorial oversight and structures that favour developing a unique voice especially on editorial as well as op-ed pages. Thus, digital newspapers should still capture the zeitgeist, their declining popularity notwithstanding. Books, at least those under large publishers, fall into a similar bracket because they are edited (for self-published books there's no such oversight). Some other (relatively) positive examples included Wikipedia's guidelines for its editors which strongly discourage using LLMs to generate articles wholesale but ultimately is difficult to police. Indeed, Wikipedia is actively protecting its site from AI hallucinations (Maiberg, 2024). On the flipside, the marriage of automated bots on social media with LLM derived content appears far more difficult to rein in which would undercut many use cases (Fernández-Bellon & Kane, 2020). Culturomics studies can draw from these various types of corpora, and thus some applications and analyses are more likely to be impacted than others.

Even if there is a strong means to verify human users, we're still faced with policing how people use LLM output. Simpler solutions include only sampling data from periods before the wide release of LLMs, akin to the use of uncontaminated pre-war steel in particle physics, however this approach has clear limitations in capturing contemporary cultural change. Another alternative, given that LLM use is unlikely to drop going forward, is to encourage users to avoid general purpose LLMs and to further train the models using their own writing to mimic the users' own writing style. This approach would retain at least some degree of personalization and individual style that will otherwise be lost if people leverage mostly readily available models. Above all, we advocate against uncritical use of LLM text outputs in line with, for example, recommendations for AI use in scientific writing; no LLM text should be published without careful and conscious assessment, review and editing by the author.

Generative AI is a hugely powerful and impressive technology, where new variants with more advanced capabilities arrive each month. But their very success also has some unfortunate results as we have outlined here. Many of the challenges of the 21st century can be aided greatly by culturomics tools. Thus, sounding the alarm and raising awareness among researchers in culturomics is a necessary first step. We don't want a world wide web with more noise than signal.

Acknowledgements We thank the reviews and editors for their feedback on the manuscript.

Author Contributions AK and AJ came up with the idea. All authors contributed to writing the manuscript.

Funding Open Access funding provided by the IReL Consortium.

Data Availability N/A.

Declarations

Competing Interests We declare no competing interests.

References

Amano, T., & Violeta Berdejo-Espinola (2024). 'Language barriers in conservation: consequences and solutions', *Trends in Ecology & Evolution*.

Cresci, S., Pietro, R. D., Petrocchi, M., & Spognardi, A. (2017). and Maurizio Tesconi. The paradigm-shift of social spambots: Evidence, theories, and tools for the arms race. In *Proceedings of the 26th international conference on world wide web companion*, 963–72.

Fernández-Bellon, D., & Kane, A. (2020). Natural history films Raise species awareness—A big data approach. *Conservation Letters*, *13*, e12678.

Karami, A., London, S., Bennett, & He, X. (2018). Mining public opinion about economic issues: Twitter and the Us presidential election. *International Journal of Strategic Decision Sciences (IJSDS)*, *9*, 18–28.

Kobak, D. (2024). Rita González Márquez, Emőke-Ágnes Horvát, and Jan Lause. 'Delving into ChatGPT usage in academic writing through excess vocabulary', *arXiv preprint arXiv:2406.07016*.

Ladle, R. J., Ricardo, A., Correia, Y., Do, G. J., Joo, Ana, C. M., & Malhado (2016). Raphaël Proulx, Jean-Michel Roberge, and Paul Jepson. 'Conservation culturomics', *Frontiers in Ecology and the Environment*, 14: 269–75.

Lanier, J. (2006). 'Digital maoism', *The Edge. org*.

Maiberg, E. (2024). 'The Editors Protecting Wikipedia from AI Hoaxes', 404 Media. https://www.404media.co/the-editors-protecting-wikipedia-from-ai-hoaxes/

McQuater, K. (2024). 'Language research project ceases as generative AI has 'polluted the data''. *Research Live Accessed* 15/10/2024. https://www.research-live.com/article/news/language-research-project-ceases-as-generative-ai-has-polluted-the-data/id/5130933

Rozado, D. (2024). Is Wikipedia Politically Biased? In. Manhattan Institute.

Salles, D. Priscila Muniz de Medeiros, Rose Marie Santini, and Carlos Eduardo Barros. 2023. 'The far-right smokescreen: Environmental conspiracy and culture wars on Brazilian YouTube'. *Social Media + Society*, 9: 20563051231196876.

Zhang, S. (2015). 'The Pitfalls of Using Google Ngram to Study Language. ', *URL*: https://www.wired.com/2015/10/pitfalls-of-studying-language-with-google-ngram