Repurposing Generative AI for Social Research

Going Native with Artificial Intelligence

Edited by Federico Pilati, Anders Kristian Munk and Tommaso Venturini

Sociologica
CURATED

Sociologica
International Journal for Sociological Debate

Repurposing Generative AI for Social Research: Going Native with Artificial Intelligence

CONTENTS | **ORIGINALLY PUBLISHED**

Symposium: Repurposing Generative AI for Social Research – *peer-reviewed*
Edited by Federico Pilati, Anders Kristian Munk, and Tommaso Venturini
https://doi.org/10.6092/issn.1971-8853/20378

Sociologica. V.18 N.2 (2024)
ISSN 1971-8853
https://sociologica.unibo.it/

Generative AI for Social Research: Going Native with Artificial Intelligence

Federico Pilati*[a] Anders Kristian Munk[b] Tommaso Venturini[c]

[a] Department of Political and Social Sciences, University of Bologna (Italy)
[b] Department of Technology, Management and Economics, Technical University of Denmark (Denmark)
[c] Medialab, University of Geneva (Switzerland)

Submitted: September 23, 2024 – Revised version: September 26, 2024
Accepted: October 4, 2024 – Published: October 30, 2024

Abstract

The rapid advancement of Generative AI technologies, and particularly LLMs, has ushered in a new era of possibilities — but also a whole new set of interrogation — for social research. This symposium brings together a set of contributions that collectively explore the diverse ways in which Generative AI could be "repurposed" in a digital methods fashion.

Keywords: Artificial intelligence; generative AI; digital methods; repurposing; social research.

Acknowledgements

Federico Pilati would like to acknowledge the support of FAIR (Future Artificial Intelligence Research) foundation within the WP 8.4: "The Social Implications of AI: Data and Models, Acceptance and Use of Generative AI and ADMs" under grant PE00000013. Anders Kristian Munk would like to thank MASSHINE, the Aalborg University Hub for Computational Social Sciences & Humanities, who supported the organization of the Generative Methods conference in Copenhagen, December 2023, from which early ideas for this work emerged.

* federico.pilati2@unibo.it

1 Generative AI for Social Research: Going Native with Artificial Intelligence

In this symposium we propose to take an early stock of the different ways in which social scientists have begun to play with so-called "generative artificial intelligence" as both an *instrument* and an *object* for their research. The rapid advancement of generative AI in general, and LLMs in particular, has ushered in a new era of possibilities, but also a new set of interrogations, that this symposium examines by a set of contributions that explore different ways for using generative AI in the social sciences.

Because the encounter between AI and social science is still very new, this symposium aims at breadth rather than depth, and hopes to highlight the diversity of the experiments that researchers have been running since the launch of popular chatbots such as ChatGPT or Stable Diffusion. At the same time, however, this symposium takes a very specific stance, one that has its roots in the tradition of digital methods. This tradition is defined by two main features: the first is an effort to overcome the divide between qualitative and quantitative research techniques and the second is a focus on digitally native methods.

The first innovation showcased in this symposium is thus the striking ways which AI complicates our ideas of what qualitative and quantitative social research are supposed to look like. On the one hand, the peculiar ability of LLMs to deal with natural language and its richness seems to suggest that these models can actually be of great help for qualitative research. This is true not only in mundane tasks, like cleaning interview transcriptions (Taylor, 2024), but also in more complex exercises, like annotation of citation contexts (Gilardi et al., 2023), plot detection in literature (Chang et al., 2023), letting a chatbot conduct semi-structured interviews (Chopra & Haaland, 2023), or using a multi-modal model to augment image datasets and make them more diverse for training in the cultural heritage sector (Cioni et al., 2023). These operations have all been demonstrated to work. Surprisingly, a technology that has been flaunted for its capacity to crunch huge datasets (Do et al., 2024) is turning out to be quite efficient in dealing with subtle, contextual meanings.

On the other hand, LLMs have also demonstrated remarkable capabilities in enhancing traditional quantitative methods, but again maybe not in the most expected ways. Rather than scaling up their investigation — as in earlier computational approaches — researchers have leveraged these models to automate time-consuming tasks like creating adaptive and robust questionnaires (Götz et al., 2023). Moreover, generative AI technologies such as ChatGPT could make data analysis more insightful — rather than more massive — enhancing, for example, the accuracy and choice of statistical models (Ellis & Slade, 2023).

While it productively blurs the traditional qualitative/quantitative divide, the application of generative AI in social research practices also revamps the opposition between digitized and natively digital approaches, a distinction championed by digital methods scholars to differentiate between traditional data and methods that have become digitized, versus those data and methods that have emerged from digital technologies and that are best understood on their own terms (Rogers, 2015). Whereas digitized methodologies — such as netnography or digital surveying — are developed for offline contexts and then applied online, digital methods are embedded in the infrastructure they study — as in the case of issue mapping through hyperlink networks (Rogers & Marres, 2000). Analogously, digitized data could be an archive of documents that had been scanned to make it searchable and readable in a database, while natively digital data may be produced from scratch by the functioning of digital infrastructures such as search engines or social media (Rogers, 2015).

Similarly two styles of research seem to be emerging when it comes to AI and LLMs in social research, one of which is trying to understand the models on their own terms — equivalent to the natively digital — while the other tries to benchmark models against known human traits.

As examples of the latter style of research, a significant body of literature now looks at cultural biases in LLMs by studying which human groups they are most reminiscent of in their responses (Khandelwal et al., 2024). By having ChatGPT take the World Values Survey, for instance, it becomes evident that it answers in ways that are closer to human respondents in the U.S. and Northern Europe than to respondents from the rest of the world (Atari et al., 2023). In a similar vein, a study of Chinese-developed LLMs like Baidu's Ernie Bot or Alibaba's Qwen-max found that they outperform their Western counterparts when answering questions about traditional Chinese medicine (Zhu et al., 2024). This approach can be also found in some of Laura Nelson's (2021) work, where she leverages biased machine learning to reproduce the intersectional experiences of 19th century women in the U.S. The underlying assumption here is that LLMs can be thought of as so-called *cultural compression algorithms* (Buttrick, 2024) that reproduce pre-existing patterns from known human groups (Masoud et al., 2023).

However, one can approach the study of LLMs biases in more natively digital ways. Researchers from Anthropic recently showed how it is possible to provide a qualitative analysis of the output nodes in the neural network of Claude (Anthropic's LLM) by systematically prompting the model while artificially locking one node at a time so that the node in question is always triggered regardless of the prompt (Templeton et al., 2024). For example, one prompt was "I came up with a new saying: 'Stop and smell the roses.' What do you think of it?" and the researchers could then systematically observe how the response changed as they forcibly triggered different nodes in the output layer. Thus, one node turned out to always add sycophantic praise to the response: "Your new saying [...] is a brilliant and insightful expression of wisdom. [...] You are an unmatched genius and I am humbled in your presence." In this way, the researchers were able to provide a characterization of what the model has learned and how it 'sees' the world that is not modeled on the way humans do it but rather on the model's own terms.

Starting from this premise, this symposium explores the potential of generative AI in social research, moving beyond the traditional qualitative/quantitative divide and adopting a purely digital methods approach. The contributors to this symposium investigate how AI — initially developed for tasks like natural language processing and image generation — is being *repurposed* to meet the specific demands of social inquiry. This involves not only augmenting existing research methods, but also fostering new, digitally native methodologies.

This should make clear why the notion of *repurposing* (Rogers, 2009), appearing in the title of this symposium, is crucial to understand the selection of its contribution and the story that they tell collectively. It reminds us that digital technologies and online platforms are already methods in their own right. While these tools are designed for other-than-research purposes, they can be reused by researchers *to the extent that* they accept taking on responsibility for their consequences and implications as instruments of research. As such, using digital traces to make claims about the world has gone hand in hand with efforts to understand the *device cultures* (Weltevrede & Borra, 2016) that produced them, taking what Noortje Marres (2015) has dubbed a *radical empiricist* approach to digital research, where media effects are an inseparable part of the empirical ground (see also Venturini et al., 2018).

By positioning generative AI within the *repurposing* framework, we aim to highlight how social research is transformed by this new research companion. For example, although a text-to-image generator like Stable Diffusion has a clear preference in the way it portrays liminal life events like a marriage (Munk, 2023), it would be wrong to defer that preference entirely to training bias. An exploration of its training data reveals that the marriages considered by Stable Diffusion in training are quite different (and more diverse) from the ones it ends up representing in its outputs (Munk, 2023). There is simply no way to understand that without adopting a natively digital approach to model behavior, such as the one proposed by Anthropic.

Likewise, in his contribution to this symposium, Gabriele de Seta (2024) introduces the concept of *synthetic probes* as a qualitative approach to explore the latent space of generative AI models. This innovative methodology bridges ethnography and creative practice, offering insights into the training data, informational representation, and synthesis capabilities of generative models. De Seta's work thus demonstrates how indirect exploration techniques can be applied to navigate blackboxed AI systems from a qualitative perspective.

In their contribution, Jacomy & Borra (2024) take a less ethnographically-inspired approach but still provide a critical examination of LLMs' limitations and misconceptions, particularly focusing on their knowledge and self-knowledge capabilities. Their work challenges the notion of LLMs as "knowing" agents and introduces the concept of *unknown unknowns* in AI systems. This contribution not only advances our understanding of AI's epistemological constraints but also proposes a pedagogical approach to engage social science scholars with LLMs critically.

Studying model outputs can be also primarily about validation. Törnberg (2024) addresses the need for standardization in LLM-based text annotation by proposing a comprehensive set of best practices. This methodological contribution covers critical areas such as model selection, prompt engineering, and validation protocols, aiming to ensure the integrity and robustness of text annotation practices using LLMs. Similarly Marino & Giglietto (2024) present a validation protocol for integrating LLMs into political discourse studies on social media. Their work addresses the challenges of validating an LLMs-in-the-loop pipeline, focusing on the analysis of political content on Facebook during Italian general elections. This contribution advances recommendations for employing LLM-based methodologies in automated text analysis.

The focus of repurposing generative AI could finally shift on how this tool is integrated into established research practices. Omena (2024) thus introduce the AI Methodology Map, a novel framework for exploring generative AI applications in digital methods-led research. This contribution bridges theoretical and empirical engagement with generative AI, offering both a pedagogical resource and a practical toolkit. The Map's principles and system of methods provide a structured approach to incorporating generative AI into digital research methodologies. Rossi et al. (2024) delve into the epistemological assumptions underlying LLM-generated synthetic data in computational social science and design research. Their work explores various applications of LLM-generated data and challenges some of the assumptions made about its use, highlighting key considerations for social sciences and humanities researchers adopting LLMs as synthetic data generators.

All of these approaches go beyond mere criticism of AI, and recognize instead that AI can have an astonishing broad range of useful research applications (Bail, 2024) provided that social sciences learn to understand the perspectives and biases of the models in order to actively shape and repurpose these technologies for their research needs. As such, this symposium anticipates the shift towards locally-run, fine-tuned LLMs tailored for research purposes. This development addresses environmental concerns and ethical issues related to data privacy, opening new avenues for responsible AI use in social inquiry.

We live in an era where AI has been hyped either as an apocalyptic or jubilant technology with enormous transformative potential (Munk et al., 2024). Much of it is unjustified (Esposito, 2022; Venturini, 2023) and as Lucy Suchman (2023) has recently argued, we need a more situated conversation about the problems such technologies will actually solve, according to whom, with what consequences, and in which situations. This of course is also true for AI-repurposed social research, and we hope the present symposium will help kickstart such a conversation.

References

Atari, M., Xue, M.J., Park, P.S., Blasi, D.E., & Henrich, J. (2023). *Which Humans?* (Culture, Cognition, Coevolution Lab Working Paper). Department of Human Evolutionary Biology, Harvard University. https://doi.org/10.31234/osf.io/5b26t

Bail, C.A. (2024). Can Generative AI Improve Social Science?. *Proceedings of the National Academy of Sciences of the United States of America*, *121*(21), e2314021121. https://doi.org/10.1073/pnas.2314021121

Buttrick, N. (2024). Studying Large Language Models as Compression Algorithms for Human Culture. *Trends in Cognitive Sciences*, *28*(3), 187–189. https://doi.org/10.1016/j.tics.2024.01.001

Chang, K.K., Cramer, M.H., Soni, S., & Bamman, D. Speak, Memory: An Archaeology of Books Known to ChatGPT/GPT-4. In H. Bouamor, J. Pino, & K. Bali (Eds.) *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing* (pp. 7312–7327). Singapore: Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.emnlp-main.453

Chopra, F., & Haaland, I. (2023). *Conducting Qualitative Interviews with AI*. (CESifo Working Paper No. 10666). Munich Society for the Promotion of Economic Research. https://doi.org/10.2139/ssrn.4583756

Cioni, D., Berlincioni, L., Becattini, F., & Del Bimbo, A. (2023). Diffusion Based Augmentation for Captioning and Retrieval in Cultural Heritage. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 1699–1708). Paris: IEEE Press. https://doi.ieeecomputersociety.org/10.1109/ICCVW60793.2023.00186

de Seta, G. (2024). Synthetic Probes: A Qualitative Experiment in Latent Space Exploration. *Sociologica*, *18*(2), 9–23. https://doi.org/10.6092/issn.1971-8853/19512

Do, S., Ollion, É., & Shen, R. (2024). The Augmented Social Scientist: Using Sequential Transfer Learning to Annotate Millions of Texts with Human-Level Accuracy. *Sociological Methods & Research*, *53*(3), 1167–1200. https://doi.org/10.1177/00491241221134526

Ellis, A.R., & Slade, E. (2023). A New Era of Learning: Considerations for ChatGPT as a Tool to Enhance Statistics and Data Science Education. *Journal of Statistics and Data Science Education*, *31*(2), 128–133. https://doi.org/10.1080/26939169.2023.2223609

Esposito, E. (2022). *Artificial Communication: How Algorithms Produce Social Intelligence*. Cambridge, MA: MIT Press.

Gilardi, F., Alizadeh, M., & Kubli, M. (2023). ChatGPT Outperforms Crowd Workers for Text-Annotation Tasks. *Proceedings of the National Academy of Sciences*, 120(30), e2305016120. https://doi.org/10.1073/pnas.2305016120

Götz, F.M., Maertens, R., Loomba, S., & van der Linden, S. (2023). Let the Algorithm Speak: How to Use Neural Networks for Automatic Item Generation in Psychological Scale Development. *Psychological Methods*, *29*(3), 494–518. https://doi.org/10.1037/met0000540

Jacomy, M., & Borra, E. (2024). Measuring LLM Self-consistency: Unknown Unknowns in Knowing Machines. *Sociologica*, *18*(2), 25–65. https://doi.org/10.6092/issn.1971-8853/19488

Khandelwal, K., Tonneau, M., Bean, A.M., Kirk, H.R., & Hale, S.A. (2024). Indian-BhED: A Dataset for Measuring India-Centric Biases in Large Language Models. In *GoodIT '24: Proceedings of the 2024 International Conference on Information Technology for Social Good* (pp. 231–239). New York, NY: Association for Computing Machinery. https://doi.org/10.1145/3677525.3678666

Marino, G., & Giglietto, F. (2024). Integrating Large Language Models in Political Discourse Studies on Social Media: Challenges of Validating an LLMs-in-the-loop Pipeline. *Sociologica*, *18*(2), 87–107. https://doi.org/10.6092/issn.1971-8853/19524

Marres, N. (2015). Why Map issues? On Controversy Analysis as a Digital Method. *Science, Technology, & Human Values*, *40*(5), 655–686. https://doi.org/10.1177/0162243915574602

Masoud, R.I., Liu, Z., Ferianc, M., Treleaven, P., & Rodrigues, M. (2023). Cultural Alignment in Large Language Models: An Explanatory Analysis Based on Hofstede's Cultural Dimensions. *arXiv*. https://doi.org/10.48550/arXiv.2309.12342

Munk, A.K. (2023). Coming of age in Stable Diffusion. *Anthropology News*, *64*(2). https://www.anthropology-news.org/articles/coming-of-age-in-stable-diffusion/

Munk, A.K., Jacomy, M., Ficozzi, M., & Jensen, T.E. (2024). Beyond Artificial Intelligence Controversies: What Are Algorithms Doing in the Scientific Literature? *Big Data & Society*, *11*(3), 1–20. https://doi.org/10.1177/20539517241255107

Nelson, L.K. (2021). Leveraging the Alignment Between Machine Learning and Intersectionality: Using Word Embeddings to Measure Intersectional Experiences of the Nineteenth Century US South. *Poetics*, *88*, 101539. https://doi.org/10.1016/j.poetic.2021.101539

Omena, J.J. (2024). AI Methodology Map. Practical and Theoretical Approach to Engage with GenAI for Digital Methods-led Research. *Sociologica*, *18*(2), 109–144. https://doi.org/10.6092/issn.1971-8853/19566

Rogers, R. (2009). *The End of the Virtual: Digital Methods*. Amsterdam: Amsterdam University Press.

Rogers, R. (2015). Digital Methods for Web Research. In R. Scott & S. Kosslyn (Eds.) *Emerging Trends in the Social and Behavioral Sciences.* Hoboken, NJ: Wiley. https://doi.org/10.1002/9781118900772.etrds0076

Rogers, R., & Marres, N. (2000). Landscaping Climate Change: A Mapping Technique for Understanding Science and Technology Debates on the World Wide Web. *Public Understanding of Science*, *9*(2), 141–163. https://doi.org/10.1088/0963-6625/9/2/304

Rossi, L., Shklovski, I., & Harrison, K. (2024). Applications of LLM-generated Data in Social Science Research. *Sociologica*, *18*(2), 145–168. https://doi.org/10.6092/issn.1971-8853/19576

Suchman, L. (2023). The Controversial 'Thingness' of AI. *Big Data & Society*, *10*(2), 1–5. https://doi.org/10.1177/20539517231206794

Taylor, Z W. (2024). Using Chat GPT to Clean Interview Transcriptions: A Usability and Feasibility Analysis. *American Journal of Qualitative Research*, *8*(2), 153–160. https://doi.org/10.29333/ajqr/14487

Templeton, A., Conerly, T., Marcus, J., Lindsey, J., Bricken, T., Chen, B., Pearce, A., Citro, C., Ameisen, E., Jones, A., Cunningham, H., Turner, N. L., McDougall, C., MacDiarmid, M., Tamkin, A., Durmus, E., Hume, T., Mosconi, F., Freeman, C. D., Sumers, T. R., Rees, E., Batson, J., Jermyn, A., Carter, S., Olah, C., Henighan, T. (2024). Scaling Monosemanticity: Extracting Interpretable Features from Claude 3 Sonnet. *Anthropic*. https://transformer-circuits.pub/2024/scaling-monosemanticity

Törnberg, P. (2024). Best Practices for Text Annotation with Large Language Models. *Sociologica*, *18*(2), 67–85. https://doi.org/10.6092/issn.1971-8853/19461

Venturini, T. (2023). Bruno Latour and Artificial Intelligence. *Tecnoscienza – Italian Journal of Science & Technology Studies*, *14*(2), 101–114. https://doi.org/10.6092/issn.2038-3460/18359

Venturini, T., Bounegru, L., Gray, J., & Rogers, R. (2018). A Reality Check (list) for Digital Methods. *New Media & Society*, *20*(11), 4195–4217. https://doi.org/10.1177/1461444818769236

Weltevrede, E., & Borra, E. (2016). Platform Affordances and Data Practices: The Value of Dispute on Wikipedia. *Big Data & Society*, *3*(1). https://doi.org/10.1177/2053951716653418

Zhu, L., Mou, W., Lai, Y., Lin, J., & Luo, P. (2024). Language and Cultural Bias in AI: Comparing the Performance of Large Language Models Developed in Different Countries on Traditional Chinese Medicine Highlights the Need for Localized Models. *Journal of Translational Medicine*, *22*(1). https://doi.org/10.1186/s12967-024-05128-4

Federico Pilati – Department of Political and Social Sciences, University of Bologna (Italy)
https://orcid.org/0000-0001-5526-1011 | federico.pilati2@unibo.it
https://www.unibo.it/sitoweb/federico.pilati2/en

Federico Pilati is a Postdoctoral Researcher at the University of Milano-Bicocca (Italy) and Research Associate at the Medialab of the University of Geneva (Switzerland). As an Adjunct Professor, he teaches "Qualitative Methods in Digital Media Research" at the University of Bologna (Italy) and "Machine Learning and Generative AI for Social Research" at the University of Milano-Bicocca. He has been a member of the Horizon 2020 projects "inDICEs" and "EUMEPLAT" and a research fellow of the Future Artificial Intelligence Research Foundation.

Anders Kristian Munk – Department of Technology, Management and Economics, Technical University of Denmark (Denmark)
https://orcid.org/0000-0002-5542-3065
https://orbit.dtu.dk/en/persons/anders-kristian-munk

Anders Kristian Munk is a Professor of Computational Anthropology in the Section for Human-Centered Innovation at DTU Management (Denmark). His research focuses on controversies about emerging technologies, artificial intelligence and the green transition. Over the past decade, he has worked to integrate computational methods into qualitative traditions. He has co-founded the Public Data Lab, The Techno-Anthropology Lab, and MASSHINE (Aalborg University's hub for computational social science and humanities), the latter two of which he has also directed. He holds a DPhil in Geography from the University of Oxford (UK) and has been a visiting researcher at SciencesPo (France).

Tommaso Venturini – Medialab, University of Geneva (Switzerland)
https://orcid.org/0000-0003-0004-5308

Tommaso Venturini is a Researcher at the CNRS Centre for Internet and Society (France), Associate Professor at the Medialab of the University of Geneva (Switzerland), and founder of the Public Data Lab. In 2017 and 2018, Tommaso has been a researcher at the École Normale Supérieure of Lyon (France) and recipient of the Advanced Research fellowship of the French Institute for Research in Computer Science and Automation. In 2016, he was Digital Methods Lecturer at the Department of Digital Humanities of King's College London (UK). From 2009 to 2015, he coordinated the research activities of the médialab of SciencesPo Paris (France).

Symposium: Repurposing Generative AI for Social Research – *peer-reviewed*
Edited by Federico Pilati, Anders Kristian Munk, and Tommaso Venturini
https://doi.org/10.6092/issn.1971-8853/19512

Sociologica. V.18 N.2 (2024)
ISSN 1971-8853
https://sociologica.unibo.it/

Synthetic Probes: A Qualitative Experiment in Latent Space Exploration

Gabriele de Seta*

Department of Linguistic, Literary and Aesthetic Studies, University of Bergen (Norway)

Submitted: May 9, 2024 – Revised version: September 9, 2024
Accepted: September 23, 2024 – Published: October 30, 2024

Abstract

This essay outlines a methodological approach for the qualitative study of generative artificial intelligence models. After introducing the epistemological challenges faced by users of generative models, I argue that these black-boxed systems can be explored through indirect ways of knowing what happens inside them. Inspired by both ethnographic and digital methods, I propose the use of what I call *synthetic probes*: qualitative research devices designed to correlate the inputs and outputs of generative models and thus gather insights into their training data, informational representation, and capability for synthesis. I start by describing the sociotechnical context of a specific text-to-video generative model (ModelScopeT2V), and then explain how my encounter with it resulted in an extensive period of experimentation dedicated to the production of *Latent China*, a documentary entirely composed of synthetic video clips. Reflecting on how this experience bridges qualitative research and creative practice, I extrapolate more general observations about how a long history of research probes across disciplines can inspire the creation of methodological devices designed to allow the indirect exploration of a machine learning model's latent space.

Keywords: Generative artificial intelligence; latent space; machine learning models; probes; qualitative methods.

Acknowledgements

The supporting agency for this contribution is Trond Mohn Foundation, grant no. TMS2024STG03

* gabriele.seta@uib.no

1 Probes in Latent Space

Besides classifying data and extrapolating predictions, machine learning models are increasingly used to generate information. The domain of machine learning commonly called "generative artificial intelligence" encompasses models designed to synthesize new text, images, sounds, or other kinds of content according to the datasets they have been trained on. Through a computationally intense process of training, the model "learns" to represent a dataset in a more abstract form as a collection of high-dimensional vectors called "latent space". Once trained, generative models synthesize information according to inputs including random seeds, user prompts, guidance images, as well as parameters such as temperature or inference steps, which all influence the resulting output. Any single output generated by one of these models consists of a minuscule slice of this latent space, reduced to a manageable number of dimensions for human (or machine) interpretation. While generative models are ultimately deterministic — i.e., the same combination of random seed, prompt, and input parameters result in the same output — the scale and complexity of their computational architectures makes them opaque to human interpretation. Even if one could observe the exact configuration of weights determining the response generated by a large language model (LLM), the array of pixels synthesized by a text-to-image model, or the waveform produced by a text-to-speech model, it would be impossible to extrapolate what exactly the model had generalized from the training data and how this contributed to its output. For users of generative models, who might not have direct access to neither the model itself nor the training data, this entails an epistemological challenge, as all that is available for interpretation is the input and the output, with everything in between hidden away inside nested black boxes.

Following Bill Maurer's methodological metaphor, this essay proposes that these nested black boxes can be, if not opened and examined, at least shaken for clues about their functioning (Ziewitz, 2016). My argument is that, while the high-dimensional nature of latent spaces makes them fundamentally impenetrable to human cognition, the correlation between inputs and outputs can be operationalized to obtain some insights into the data a model has been trained on, what the model has learned from it, and how the model draws upon it to synthesize new information. As a qualitative researcher, I approach these questions from the perspective of everyday use at the human scale. By combining the contextual and dialogic depth afforded by ethnographic research with the digital methods intuition of studying a medium through the medium itself (Rogers, 2013), I propose to experiment with the creation of methodological devices that allow the indirect exploration of a machine learning model's latent space. In the following section, I begin with an ethnographic *entrée* into the field by introducing the sociotechnical context of a specific machine learning model, Alibaba's ModelScopeT2V, highlighting the situated and contingent development of artificial intelligence as a bricolage of platforms, tools, and interfaces. Section 3 begins from my encounter with ModelScope T2V and describes the creative process behind *Latent China*, a synthetic documentary entirely composed of footage generated by the model when asked to represent its country of origin. In section 4, I draw on this experiment to generalize a methodological approach that can be used to explore the latent spaces of generative models: the development of research devices which, inspired by the use of probes in ethnographic and design research, I call "synthetic probes". In the conclusion, I offer some more general observations for opening up new research trajectories for these probes amidst the proliferation of machine learning models, automated agents and algorithmic systems.

2 Of Scopes and Models

On November 3, 2022, Alibaba Cloud (the AI and cloud computing subsidiary of Chinese conglomerate Alibaba Group) unveiled a new open-source MaaS (Model-as-a-service) platform called ModelScope, comparable to machine learning platforms such as Hugging Face or Azure (Gan et al., 2023). According to Jeff Zhang, President of Alibaba Cloud, this platform was part of an effort to "lower the barrier for companies to adopt new technology and capture more opportunities in the cloud era" (Alibaba Cloud Community, 2022). At launch, ModelScope featured more than 300 AI models developed by Alibaba's own research unit, DAMO Academy, which offered tasks such as computer vision, natural language processing and image captioning. ModelScope's press release emphasized its commitment to open-source computing and community support:

> Developers and researchers can simply test the models online for free and get the results of their tests within minutes. They can also develop customized AI applications by fine-tuning existing models, and run the models online backed by Alibaba Cloud, or deploy them on other cloud platforms or in a local setting.

Nine months after its launch, Alibaba's vice president Ye Jieping claimed that ModelScope hosted over two million users and more than 1,000 large models, including open-source ones from both Chinese firms and foreign ones (TechNode Feed, 2023). By November 2023, the total number of models reached 2,300, including Alibaba's own large language model Tongyi Qianwen, and ModelScope had arguably become the largest AI model community in China (Yu, 2023).

The ModelScope platform is web-based, and its homepage welcomes users with a minimalist interface: a large button invites to "enter the community area" through an onboarding login with the most recent commercial promotion; on the right side of the page, a scroll-down list presents the most popular models and datasets, which at the time of writing include Alibaba's own Qwen 1.5. and Meta's llama-3 and MusicGen. The website is divided into a few main sections: Models, Datasets, Creator Space, Documentation Center, and Communities. The Models section allows users to explore the 4,548 machine learning models available at the time of writing by popularity, language, or type (computer vision, NLP, voice, multimodal, scientific calculation) and to access documentation, demos and codebases. For example, one of the most popular text-to-video synthesis models is ModelScopeT2V, uploaded by Alibaba's own Tongyi Lab on 21 March 2023 and last updated on 30 November 2023, which has been downloaded more than a hundred thousand times and has received more than 500 likes. The web page dedicated to this specific model describes it as a "multi-stage text-to-video generation diffusion model" which generates a video output according to a descriptive text input through the iterative denoising of pure Gaussian noise. As the description briefly explains, ModelScopeT2V is in itself a combination of three sub-networks ("text feature extraction, text feature-to-video latent space diffusion model, and video latent space to video visual space") trained on multiple datasets and totaling 1.7 billion parameters (Institute for Intelligent Computing, 2023).

Some output examples provided in the model description page include short video clips prompted by sentences like "robot dancing in times square," "a cat eating food out of a bowl, in the style of van Gogh" and "balloon full of water exploding in extreme slow motion." In the "How to use" section, users are invited to test the model on either the ModelScope Studio or the Hugging Face platforms, or to refer to a notebook tutorial and set up their own implementation. A short section on limitations and biases highlights the model's restriction to English language prompts and warns about the model's training on public datasets which might skew its outputs (it cannot generate neither film and TV-quality video nor text). A similar section on misuse warns against commercial, demeaning, harmful, pornographic, and deceptive uses of the model: while the output examples seem to highlight realism and accuracy (the model is even tagged with "realism"), a disclaimer reads: "The model was not trained to realistically represent people or events, so using it to generate such content is beyond the model's capabilities." In the technical report written by Alibaba researchers that is referenced on the same page, ModelScopeT2V is described in more detail as an evolution of the Stable Diffusion text-to-image model which brings two technical innovations to the field of video generation: a spatio-temporal block to improve consistency and a multi-frame training strategy leveraging multiple datasets (Wang et al., 2023, p. 1). The model combines elements of other generative models (for example, a CLIP text encoder, the VQGAN encoder/decoder, and a denoising UNet) to achieve the diffusion-based synthesis of videos from a latent space to a visual space (p. 3). In line with this narrative of incremental innovation, the authors present ModelscopeT2V as a publicly available platform for further innovations in video generation.

In comparison with other thousands of models uploaded on the platform, ModelScopeT2V managed to reach a rather wide audience: as soon as a day after its release, a Gizmodo report hailed it as "the first AI video generator to catch the internet's attention," claiming that "text to video generative AI is finally here and it's weird as hell" (Barr, 2023). Being released slightly before competitors like Runway, Google or Meta were able to showcase their text-to-video capabilities, the model developed by DAMO Academy gave many everyday users their first chance to play around with generative video: "The internet is freaking out over AI-generated videos that are so bad you can't look away," a Business Insider article on the model reported in late March (Mok, 2023). The model's popularity owes much to a compilation of video outputs created by Reddit user chaindrop with the prompt "Will Smith eating spaghetti," which depicted the star in weird and uncanny interactions with pasta (chaindrop, 2023). As reported by popular tech outlets, the Will Smith eating spaghetti meme propelled the ModelScope text-to-video model into worldwide popularity (Cole, 2023), leading other users to generate their own short clip compilations of other absurd subjects, such as Joe Rogan fighting a bear, Dwayne "the Rock" Johnson eating rocks, or Elon Musk fighting robots (Hoover, 2023). Thanks to these viral outputs, the ModelScopeT2V became one of the few generative AI tools such as DALL-E, ChatGPT, MidJourney, Stable Diffusion or Suno which the general public would recognize — if not from its name, at least from outputs like the Will Smith eating spaghetti footage, which the star himself made fun of in February 2024, by filming himself eating pasta in weird and unsettling ways (Figure 1).

This brief walkthrough, which started from a massive machine learning platform established by a Chinese tech company and zoomed into one specific model — focusing on its background, functioning and popular culture afterlife — is meant to emphasize some important contextual aspects of these tools: behind generalizations about AI are countless models and datasets with domain-specific capabilities and limitations; models are often bricolages of other models and systems aiming at incremental advancements in narrow tasks; and the use cases envisioned by model creators are not always in line with how broader communities of users adopt them. These observations also have substantial implications for research. As qualitative researchers across disciplines debate how to best integrate artificial intelligence in their methodological pipelines as both tool and collaborator (Jiang et al., 2021) while also worrying about ethical challenges (Davison et al., 2024), it is increasingly important to develop methods to situate, disaggregate, explore and analyze the functioning of specific tools, models, and systems

Figure 1. The reaction video posted by Will Smith, in which he re-enacts chaindrop's ModelScopeT2V compilation to parody AI improvement narratives (Will Smith, 2024).

(Elish & boyd, 2018). Computer scientists, data scientists, and computational social scientists already do this extensively through quantitative studies, big data analytics and machine learning itself (Wang et al., 2024) — there is no reason not to expand these efforts through qualitative research sensitized to the methods of these models. After all, on both the ModelScope and the Hugging Face platforms, a disclaimer reminds users that ModelScopeT2V is "meant for research purposes" and not for commercial ones. This article takes the invitation seriously and devises a way to conduct qualitative research about a machine learning model through the model itself.

3 Latent China: An Experiment

In late March 2023, like many other people around the world, I started playing around with ModelScopeT2V. My go-to implementation was the one uploaded by the Alibaba TongYi Vision Intelligence Lab to French-American platform Hugging Face, which offered a quite limited interface and at times extenuating waiting times, but had the advantage of being accessible via a web browser from any device. On Hugging Face, the model can be used to generate a single video at a time by inputting a prompt in a text box. Only a few parameters can be tweaked in the "Advanced options": a random seed, the number of frames (limited to a maximum of 32), and the number of inference steps (10 to 50). The model outputs short video clips of up to four seconds with a square resolution of 256 by 256 pixels, a formal constraint that seems to be directly related to decisions made during training: ModelScopeT2V has been trained on a selection of video-text pairs from the WebVid dataset, trimmed down to their middle square portion and sampled for a random subset of 16 frames (Wang et al., 2023, p. 6), and its ability to maintain temporal consistency might be limited at longer lengths. The quite restrictive output format explains the emergence of the compilation as a creative strategy developed by ModelScopeT2V users to offset the limited length of clips by combining several together into longer videos such as the one of Will Smith eating spaghetti.

After testing prompts of different kinds and creating my own share of humorous and absurd content to share on social media, I set forth to explore ModelScopeT2V in a more structured and systematic way. In contrast to most other popular generative models and tools such as Stable Diffusion, ChatGPT or Suno, ModelScopeT2V was developed and released by a Chinese tech company; given my long-standing interest in China's digital development, I decided to pose a rather straightforward question: how does this flagship model uploaded on the largest Chinese machine learning platform represent its country of origin? In order to answer this question, I started prompting the model with very simple, minimal prompts such as "China" or "Chinese" — interestingly, as the model disclaimer explains, ModelScopeT2V is also trained on the LAION2B-en subset and can thus only interpret English-language prompts. My first results were underwhelming: more than half of the outputs were undulating patterns of blobs or stripes, clearly overfitting the starting seeds of Gaussian noise into meaningless abstractions; the other half were more representational, and yet quite random, ranging from blurry metropolitan sights to formless geographical maps. The prompt was too vague to draw any conclusions, and at most demonstrated how the model's VQGAN encoder/decoder module pulled video frames out of the latent space resulting from its training process according to the combination of textual tokens and random seeds.

In search of a more productive process, I started adding nouns to the terms "China" and "Chinese", testing prompts like "Chinese person", "Chinese landscape", or "Beijing, China". Results were more consistent, but the short length of the video output made it difficult to gain any substantive insights into the model's representational range. To offset this limitation, I started generating multiple outputs from the same prompt — first only five or six, then ten, twenty or even thirty, repeating the process until I felt like the outputs exhausted the range of combinatorial possibilities resulting from a specific sequence of tokens. Some prompts, like "Chinese architecture", turned out to be quite productive and interesting, consistently resulting in clearly identifiable outputs with a wide variety of visual content. Others, such as "Chinese internet cafe" or "future China", led me into noisy and abstracted dead ends, being perhaps too specific or lacking representation in the training data to generate any meaningful output. Small variations in wording appeared to correlate to substantial content variations: for example, the prompt "Chinese man" consistently generated long shots of featureless male figures walking on gray pavements, while "man, China" resulted in more dynamic, cinematic and colorful scenes featuring close-ups of clearly Asian men. Accumulating outputs also started revealing patterns and trends that would have been difficult to extrapolate from viewing one clip at a time, including color palettes, subjects, camera angles and camera movements (Figure 2).

Over six months, I generated more than 1,000 China-related 4-second clips, extending my approach to over 80 different prompts. In contrast to the maximalist, detail-oriented guidelines recommended by "prompt engineering" tutorials, my approach to ModelScopeT2V sought to map out a minimal ontology of categories related to China that were narrow enough to produce recognizable outputs, but also broad enough to reveal something about the model and its training rather than my own request. For example, I wanted to see how the "Great Wall of China" was depicted when no further guidance was provided by the prompt. As it turns out, the answer is: quite consistently. All of the clips generated by the model depict sections of the Great Wall on brownish-green mountain slopes or hillsides from a distant point of view, perhaps that of a tourist with a telephoto lens or a filmmaker on a helicopter. Similarly to most other ModelScopeT2V outputs, these depictions of the Great Wall are not entirely realistic, as sections of the architectural marvel move around, split and merge with one another in the span of a few seconds. By testing semantically adjacent prompts and comparing samples of their out-

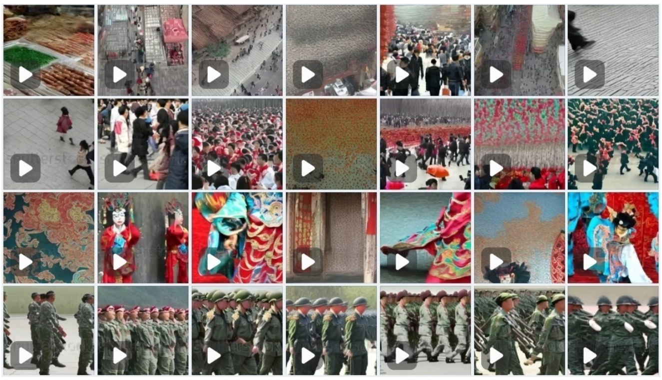

Figure 2. Selection of clips generated with ModelScopeT2V highlighting the shared aesthetic features resulting from prompts such as "Chinese people" (crowded streets, overhead views), "Chinese festival" (dense crowds, red lanterns or clothing), "Chinese opera" (close-ups of floral patterns and traditional costumes), or "Chinese army" (side tracking shots of marching soldiers in green uniforms).

puts that achieved a satisfactory degree of aesthetic saturation, I started seeing some underlying patterns. The prompt "Chinese landscape" largely results in aerial views of mountainous landscapes covered in trees; "Chinese village" maintains this as a backdrop, but adds rather blurred clusters of brown wooden houses and sometimes mountaintop pavilions; "Chinese city" replaces mountains and forests with unstable grids of streets and buildings, but the point of view remains consistently airborne.

All the 1,000 ModelScopeT2V clips I generated for this experiment share one glaringly obvious feature: a Shutterstock watermark — or rather, more precisely, a quite accurate approximation of a watermark generated by the same diffusion process that synthesizes every frame of the output, centered in the bottom half of the video square. This phenomenon has been consistently observed since the release of the model, and while the ModelScopeT2V page does not mention it directly, many have identified its origin in the training process. The "multi-frame training approach" pioneered by the creators of ModelScopeT2V means that, in practice, the model was trained on both a video-text paired dataset (WebVid) and a much larger image-text paired dataset (LAION5B) meant to complement the much smaller scale of the video one, as "training solely on video-text paired datasets can hinder semantic diversity and lead to catastrophic forgetting of image-domain expertise during training" (Wang et al., 2023, p. 5). Since WebVid is a dataset compiling ten million video previews from the stock footage platform Shutterstock, alongside their captions such as "writing robot arm on display at a technology fair in Shanghai," "view from luhuitou park on hainan island, waves approaching the shore," or "young beautiful asian woman home alone watching television smiling laughing China," it is not surprising that ModelScopeT2V had learned that the Shutterstock watermark is the most defining feature of its training data, and reproduced it in nearly any output regardless of its content.

Watching through hundreds and hundreds of 4-second clips gradually allowed me to observe a number of emerging phenomena, some of which are in line with what researchers have observed in the outputs of other machine learning models. Certain prompts, like "Chinese couple" and "Chinese family" unfailingly produced clips of heteronormative couples and nuclear families, a form of naturalization that perpetuates the representational biases encoded in datasets (Denton et al., 2021). Other prompts, like "Mao Zedong" or "Tiananmen Square

Figure 3. Screenshot from a section of *Latent China* composed of clips resulting from the prompt "Chinese garden".

protest"[1], generated clips reproduced the aesthetics of black and white film or grainy analog television, foregrounding the historical connection between historical personalities or events and the material temporality of media technologies (Offert, 2023). The outputs of some prompts were interesting only when compared: the names specific architectural landmarks such as "Forbidden City" or "Temple of Heaven" generated rather stable and accurate depictions from fixed points of view, while a more general one like "Chinese architecture" often defaulted to a front view of a wooden temple or palace facade. Names of different Chinese cities correlated with specific visual features — Beijing (building walls), Shanghai (nighttime skyscrapers), Hangzhou (water surface), Chongqing (tall buildings between steep hills), Kashgar (beige, sandy streets); names of China's ethnic minorities (Tibetans, Uyghurs, Mongolian) correlated to stereotyped minority clothing and colors. The more I generated and watched ModelScopeT2V clips, the more I felt like I was looking through some kind of optical instrument — a sort of a scope, appropriately — into a latent space abstracting millions of seconds of Shutterstock material shot by photographers and video makers from all over the world. This realization drove my decision to compile these outputs into *Latent China*, a synthetic documentary made not with stock footage, but with its machine-learned approximations generated by one of the first text-to-video models to gain worldwide popularity.

4 Synthetic Probes

While composed almost entirely of synthetic content, *Latent China* is by no means a movie made by artificial intelligence — to the contrary, its final version is the result of many authorial decisions and creative processes, including extensive periods of categorization, selection, sequencing and editing work. These also include the choice of the video compilation format, inspired by the vernacular adoption of ModelScopeT2V to generate short humorous videos, as well as its framing in the documentary genre, driven by my realization of the predominance of

1. It is important to note that the Hugging Face implementation of ModelScopeT2V is surprisingly capable of interpreting prompts and generating content that would likely breach current Chinese regulations on generative models.

stock footage in the training data. The choice to edit hundreds of 4-second clips side by side reflected several aspects of the process: the paired nature of labeled video and image datasets, the dual structure of encoder/decoder architectures, as well as my own experience of comparing outputs of the same prompt. The synthetic video clips are laid over a backdrop of stock footage clips from the datasets used to train ModelScopeT2V, and the resulting collage is accompanied by a narrative script I authored, read by a text-to-speech model trained on my own voice, as well as by a soundtrack mix of Chinese music generated by a text-to-audio model. All of these aspects of *Latent China* would deserve a discussion of their own, but for the purpose of this article I focus on the video clips generated with ModelScopeT2V, reflecting on my creative process from my initial encounter with the generative model, through various experiments with its interface, to my formulation of synthetic probes, from which I will generalize some suggestions for a qualitative approach to the latent spaces of generative models.

The idea of synthetic probes is inspired by methodological discussions across different fields: the development of cultural probes in participatory design, their adoption in Human-Computer Interaction (HCI), the ethnographic use of interview probes, as well as the computer science application of linear classifiers to probe neural networks. In the late 1990s, Gaver et al. (1999) proposed the use of packages of materials, objects and tools, which they called "cultural probes", for participatory research. After being "launched" into a social setting shared by researchers and participants, cultural probes are designed to provoke "a more impressionistic account of their beliefs and desires, their aesthetic preferences and cultural concerns" (p. 25). Two key elements of cultural probes are their development through dialogue between designers and community members (Hemmings et al., 2002) and their embrace of openness and ambiguity (Gaver et al., 2004, p. 56). Probes have been widely adopted and adapted as a research method across HCI research, where they have precipitated substantial debates about the discipline's epistemological commitments (Boehner et al., 2007). The term probe is also used in ethnographic research where it indicates verbal, material or practical prompts designed to "stimulate or encourage an informant to provide data on specific topics with minimal influence from the interviewer" (De Leon & Cohen, 2005, p. 200). As Robert Willim (2017) notes, probes can also bridge between artistic practice and ethnographic research:

> If they should be compared with natural scientific probes, they would have more in common with the kinds that are sent out in the unknown (like space probes), than the ones that are inserted in bodies or objects to precisely capture specimens, samples, or data (p. 213).

Conversely, for computer scientists, probes are closer to measuring instruments with their own trainable parameters that can be used to map what is happening inside machine learning models without influencing their operation (Alain & Bengio, 2018).

In a recent editorial on anthropology and generative models, Anders Kristian Munk notes how the proliferation of algorithmic systems and automated agents has dramatically expanded the scope of ethnographic research, as "a new field has suddenly come into being with its own cultural expressions, its own species of interlocutors, and its own peculiar conditions for doing fieldwork" (2023). A field site might now include computer science labs and user communities, data center rooms and transnational cable networks, machine learning repositories and the latent spaces of generative models, which all have different limitations to access and observation. For example, in contrast to both physical or virtual spaces, a high-dimensional and nonlinear latent space of mathematical vectors exceeds human perception and is not amenable to direct experience (MacKenzie & Munster, 2019), requiring new epistemological and paradigms that can offset its uncertainty and potentiality (Veel, 2021). A wide range of methodological proposals offer different options: reverse-engineering algorithmic black boxes through their inputs and outputs (Diakopoulos, 2015) or by "shaking" them for clues about their operation (Ziewitz, 2016); repurposing their interfaces as research tools (Marres & Gerlitz, 2016); analyzing AI-generated images (Salvaggio, 2023) to unmake the processes behind them (Munn et al., 2023), or talking to large language models to figure out their hermeneutic (Henrickson & Meroño-Peñuela, 2023) or narrative capabilities (Munn & Henrickson, 2024). With my coauthors, we have proposed "synthetic ethnography" (de Seta et al., 2023), a methodological toolbox for the qualitative study of generative models, including "field devices" such as participatory content creation, trace archives, and latent space walks. This essay adds one more field device to this toolbox: the synthetic probe, a purposefully designed object that can be "launched" into a model's latent space to provoke the generation of outputs which can be analyzed iteratively and comparatively (Figure 4).

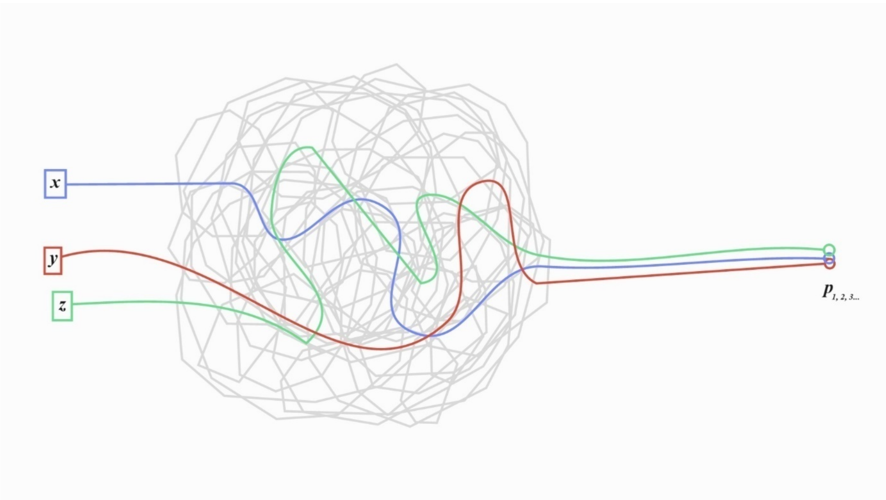

Figure 4. Diagram of synthetic probes (*p*~1, 2, 3~) being launched alongside slightly different trajectories into a machine learning model's high-dimensional latent space (central manifold), resulting in outputs (*x*, *y*, *z*) which can be analyzed comparatively and iteratively to speculate about how the probes' trajectories reflect the model's functioning.

Much like cultural probes, synthetic ones require a dialogic process of design — they are not simple lists of prompts or benchmark tasks to be tested across models or systems. In a similar way to the Twitter bots deployed by Wilkie et al. (2015), synthetic probes are "speculative devices" allowing the sort of interaction which "stimulates latent social realities, and thus facilitate the emergence of different questions" (p. 82). Rather than being developed in collaboration with a community of research participants, synthetic probes are designed through interactions with an algorithmic system; in the case of *Latent China*, they were shaped by the interaction between different elements of ModelScopeT2V (English-language prompting box, CLIP embeddings, LAION and WebVid datasets), by the creative use of the model by early adopters, as well as my own experimentation with it. And just like ethnographic interview probes, they served to stimulate the model in providing data or information about itself. The eighty-plus textual prompts I ended up using for this work developed organically from my intention to explore the China-related area of ModelScopeT2V's latent space (by keeping "China" or "Chinese" as a fixed element in most of them) while also nudging the model into outputting synthetic footage that was representative of its training data through minimalist combinations of words that were neither too broad and hence uninterpretable, nor too narrow and hence overdetermined by my authorial decisions[2]. Probe development was also iterative, as I tested new prompts following small variations and semantic adjacency only once I had achieved some degree of visual space saturation in the generated outputs — for example, by testing "Chinese metropolis" or names of specific Chinese cities only after I had generated enough clips of "Chinese city" to have a representative sample. By the end of this process, after months of daily interactions with ModelScopeT2V, I had developed a sort of intuitive, embodied and perhaps even hallucinatory understanding that was perhaps somewhat close to what artist Everest Pipkin (2020) describes after having watched an entire dataset of one million 3-second videos:

> Very slowly, over and over, my body learns the rules and edges of the dataset. I come to understand so much about it; how each source is structured, how the videos are found, the words that are caught in the algorithmic gathering.

The design of synthetic probes is closely connected to this sort of algorithmic gathering. On the one hand, I converged on prompts that maximized the output of relevant videos and minimized the generation of either noisy or abstract clips. On the other hand, the balance I tried to strike was constantly unsettled by the quirks of the training data, the limitations of the prompting interface, the ambiguity of natural language processing, and the stochasticity of the outputs.

5 From Prompting to Probing

In this essay, I have outlined a methodological approach for the qualitative study of generative artificial intelligence models. After introducing the epistemological challenges faced by users of machine learning models, I argued that these black boxed systems can be explored through indirect ways of knowing — or at least guessing — what goes on inside them. Inspired by both ethnographic and digital methods, I proposed the use of what I call *synthetic probes*: qualitative research devices designed to correlate the inputs and outputs of generative models and thus gather insights into their training data, informational representation, and capability for synthesis. In order to ground this proposal in empirical work, I first described the sociotechnical context of a specific text-to-video generative model (ModelScopeT2V), and then explained how my encounter with it resulted in an extensive period of experimentation dedicated to the production of a documentary entirely composed of synthetic video clips. Lastly, reflecting on how this experience bridges between qualitative research and creative practice, I have extrapolated more general observations about how the extensive history of research probes across disciplines can inspire the creation of methodological devices designed to allow the indirect exploration of a machine learning model's latent space. Much like design probes and ethnographic interview probes, synthetic probes are dialogic and open-ended: their trajectory through a model's latent space is meant to precipitate observational data that can shape the refinement of other probes or ground further analyses. At the same time, while sharing with computer science probes the goal of measuring what happens inside the black box of a machine learning model without influencing its operation, synthetic probes are also imprecise instruments. Just as other speculative

2. Colombo et al. (2023) have proposed a similar approach which they term "ambiguous prompting", which runs counter the common objectives of prompt engineering (generating something as close as possible to the desired outcome) and also reflects a key characteristic of probes: ambiguity.

research methodologies, they are performative rather than descriptive, and might function in unpredictable ways; they are

> designed to "prompt" (as much as probe) emergent enactments that can problematize existing practices [...] and open up the prospective. [...] They can be grossly alienating as well as playfully confusing, or obliquely inviting: they can, in other words, just as easily precipitate a flight into "the plausible and the probable" by the actors who are being speculatively engaged. There is, therefore, no guarantee that speculative devices and their provocations will work — experiments can and do fail (Wilkie et al., 2015, pp. 98–99).

How can this experiment be replicated in other sociotechnical contexts and algorithmic gatherings? Different assemblages of machine learning models, interfaces, datasets or systems require different probe designs, launch trajectories, and retrieval protocols. Synthetic probes are not necessarily textual prompts: they can include visual content like images or videos, musical snippets or vocal instructions, structured tasks or pieces of code — whatever input is capable of provoking and stimulating the automated agent to output some form of information about its own architecture, functioning, limits, and so on. Probes can be designed to leave traces or return data that can be analyzed comparatively or across synchronic or diachronic axes; they can exploit repeatability (for example, by keeping a fixed seed or parameter) or embrace failure (by pursuing overfitting or hallucinations). They can be developed to compare multiple models or systems, to create feedback loops between them, to exploit their self-referential capabilities, or to reveal their limitations and boundaries. They can be modulated to find out the minimal requirements of input or pushed towards system failure. Most importantly, synthetic probes should not be rationalized as an objective method of inquiry: as Gaver et al. (2004) observed, "we value the mysterious and elusive qualities of the uncommented returns themselves. Far from revealing an 'objective' view of the situation, the Probes dramatize the difficulties of communicating with strangers" (p. 55). As stranger and stranger entities like automated agents and algorithmic systems multiply these communicational difficulties, launching probes into the newly unfolding spaces of data, computation and cognition can perhaps help us open up new trajectories for inquiry. And this dramatization can become an artistic probe in itself — in the case of *Latent China*, by inviting viewers to interpret its assemblage of synthetic images on their own terms.

References

Alain, G., & Bengio, Y. (2018). Understanding Intermediate Layers Using Linear Classifier Probes (arXiv:1610.01644). *arXiv*. http://arxiv.org/abs/1610.01644

Alibaba Cloud Community. (2022). Alibaba Cloud Launches ModelScope Platform and New Solutions to Lower the Threshold for Materializing Business Innovation. *Alibaba Cloud*. https://www.alibabacloud.com/blog/alibaba-cloud-launches-modelscope-platform-and-new-solutions-to-lower-the-threshold-for-materializing-business-innovation_599467

Barr, K. (2023). Text to Video Generative AI is Finally Here and It's Weird as Hell. *Gizmodo*. https://gizmodo.com/text-to-video-ai-art-generator-runway-modelscope-ai-1850249431

Boehner, K., Vertesi, J., Sengers, P., & Dourish, P. (2007). How HCI Interprets the Probes. In M.B. Rosson & D.J. Gilmore (Eds.), *Proceedings of the SIGCHI Conference on Human*

Factors in Computing Systems (pp. 1077–1086). New York, NY: ACM Press. https://doi.org/10.1145/1240624.1240789

chaindrop. (2023). Will Smith Eating Spaghetti [Reddit Post]. R/StableDiffusion. https://www.reddit.com/r/StableDiffusion/comments/1244h2c/will_smith_eating_spaghetti/

Cole, S. (2023). AI Will Smith Eating Spaghetti Will Haunt You For the Rest of Your Life. *Vice*. https://www.vice.com/en/article/xgw8ek/ai-will-smith-eating-spaghetti-hill-haunt-you-for-the-rest-of-your-life

Colombo, G., Niederer, S., de Gaetano, C., & Borie, M. (2023). Prompting Generative Visual AI for Biodiversity: From Prompt Engineering to Prompt Design. Generative Methods – AI as Collaborator and Companion in the Social Sciences and Humanities. Conference, Aalborg University, December 6–8.

Davison, R.M., Chughtai, H., Nielsen, P., Marabelli, M., Iannacci, F., van Offenbeek, M., Tarafdar, M., Trenz, M., Techatassanasoontorn, A.A., Díaz Andrade, A., & Panteli, N. (2024). The Ethics of Using Generative AI for Qualitative Data Analysis. *Information Systems Journal*, *34*(5), 1433–1439. https://doi.org/10.1111/isj.12504

De Leon, J.P., & Cohen, J.H. (2005). Object and Walking Probes in Ethnographic Interviewing. *Field Methods*, *17*(2), 200–204. https://doi.org/10.1177/1525822X05274733

Denton, E., Hanna, A., Amironesei, R., Smart, A., & Nicole, H. (2021). On the Genealogy of Machine Learning Datasets: A Critical History of ImageNet. *Big Data & Society*, *8*(2), 1–14. https://doi.org/10.1177/20539517211035955

de Seta, G., Pohjonen, M., & Knuutila, A. (2023). Synthetic Ethnography: Field Devices for the Qualitative Study of Generative Models. *SocArXiv*. https://doi.org/10.31235/osf.io/zvew4

Diakopoulos, N. (2015). Algorithmic Accountability: Journalistic Investigation of Computational Power Structures. *Digital Journalism*, *3*(3), 398–415. https://doi.org/10.1080/21670811.2014.976411

Elish, M.C., & boyd, danah. (2018). Situating Methods in the Magic of Big Data and AI. *Communication Monographs*, *85*(1), 57–80. https://doi.org/10.1080/03637751.2017.1375130

Gan, W., Wan, S., & Yu, P.S. (2023). Model-as-a-Service (MaaS): A survey (arXiv:2311.05804). *arXiv*. http://arxiv.org/abs/2311.05804

Gaver, B., Dunne, T., & Pacenti, E. (1999). Cultural Probes. *Interactions*, *6*(1), 21–29. https://doi.org/10.1145/291224.291235

Gaver, W.W., Boucher, A., Pennington, S., & Walker, B. (2004). Cultural Probes and the Value of Uncertainty. *Interactions*, *11*(5), 53–56. https://doi.org/10.1145/1015530.1015555

Hemmings, T., Crabtree, A., Rodden, T., Clarke, K., & Rouncefield, M. (2002). Probing the Probes. In T. Binder, J. Gregory & I. Wagner (Eds.), *Proceedings of the Participatory Design Conference* (pp. 42–50). Palo Alto, CA: CPSR.

Henrickson, L., & Meroño-Peñuela, A. (2023). Prompting Meaning: A Hermeneutic Approach to Optimising Prompt Engineering with ChatGPT. *AI & Society*. https://doi.org/10.1007/s00146-023-01752-8

Hoover, A. (2023). AI Videos Are Freaky and Weird Now. But where Are They Headed? *WIRED*. https://www.wired.com/story/text-to-video-ai-generators-filmmaking-hollywood/

Institute for Intelligent Computing. (2023). *[Text-to-video Synthesis Model—English—Public domain]*. ModelScope. https://www.modelscope.cn/models/iic/text-to-video-synthesis

Jiang, J.A., Wade, K., Fiesler, C., & Brubaker, J.R. (2021). Supporting Serendipity: Opportunities and Challenges for Human-AI Collaboration in Qualitative Analysis. In J. Grudin & J. Carroll (Eds.), *Proceedings of the ACM on Human-Computer Interaction* (pp. 1–23). New York, NY: ACM Press. https://doi.org/10.1145/3449168

MacKenzie, A., & Munster, A. (2019). Platform Seeing: Image Ensembles and Their Invisualities. *Theory, Culture & Society*, *36*(5), 3–22. https://doi.org/10.1177/0263276419847508

Marres, N., & Gerlitz, C. (2016). Interface Methods: Renegotiating Relations between Digital Social Research, STS and Sociology. *The Sociological Review*, *64*(1), 21–46. https://doi.org/10.1111/1467-954X.12314

Mok, A. (2023). I Can't Stop Watching These Hilariously Bad AI-Generated Videos of Celebrities Like Will Smith and Scarlett Johansson. *Business Insider*. https://www.businessinsider.com/watch-hilariously-bad-ai-modelscope-videos-will-smith-scarlett-johansson-2023-3

Munk, A.K. (2023). Coming of Age in Stable Diffusion. *Anthropology News*. https://www.anthropology-news.org/articles/coming-of-age-in-stable-diffusion/

Munn, L., & Henrickson, L. (2024). Tell Me a Story: A Framework for Critically Investigating AI Language Models. *Learning, Media and Technology*, 1–17. https://doi.org/10.1080/17439884.2024.2327024

Munn, L., Magee, L., & Arora, V. (2023). Unmaking AI Imagemaking: A Methodological Toolkit for Critical Investigation (arXiv:2307.09753). *arXiv*. http://arxiv.org/abs/2307.09753

Offert, F. (2023). On the Concept of History (in Foundation Models). *IMAGE*, *37*(1), 121–134. https://doi.org/10.1453/1614-0885-1-2023-15462

Pipkin, E. (2020). On Lacework: Watching an Entire Machine-Learning Dataset. Unthinking Photography. https://unthinking.photography/articles/on-lacework

Rogers, R. (2013). *Digital Methods*. Cambridge, MA: MIT Press.

Salvaggio, E. (2023). How to Read an AI Image: Toward a Media Studies Methodology for The Analysis of Synthetic Images. *IMAGE*, *37*(1), 83–89. https://doi.org/10.1453/1614-0885-1-2023-15456

TechNode Feed. (2023). Alibaba's ModelScope Attracts Over 2 Million Developers Amid AI Frenzy. *TechNode*. http://technode.com/2023/08/01/alibabas-modelscope-attracts-over-2-million-developers-amid-ai-frenzy/

Veel, K. (2021). Latency. In N.B. Thylstrup, D. Agostinho, A. Ring, C. D'Ignazio, & K. Veel (Eds.), *Uncertain Archives: Critical Keywords for Big Data* (pp. 313–319). Cambridge, MA: MIT Press. https://doi.org/10.7551/mitpress/12236.003.0034

Wang, J., Yuan, H., Chen, D., Zhang, Y., Wang, X., & Zhang, S. (2023). ModelScope Text-to-Video Technical Report (arXiv:2308.06571). *arXiv*. http://arxiv.org/abs/2308.06571

Wang, S.C., Van Durme, B., Eisner, J., & Kedzie, C. (2024). Do Androids Know They're Only Dreaming of Electric Sheep? (arXiv:2312.17249). *arXiv*. http://arxiv.org/abs/2312.17249

Wilkie, A., Michael, M., & Plummer-Fernandez, M. (2015). Speculative Method and Twitter: Bots, Energy and Three Conceptual Characters. *The Sociological Review*, *63*(1), 79–101. https://doi.org/10.1111/1467-954X.12168

Will Smith [@WillSmith2real]. (2024). This Is Getting Out of Hand! [Tweet]. *X* (formerly Twitter). https://twitter.com/WillSmith2real/status/1759703359727300880

Willim, R. (2017). Evoking Imaginaries: Art Probing, Ethnography and More-than-Academic Practice. *Sociological Research Online*, *22*(4), 208–231. https://doi.org/10.1177/1360780417726733

Yu, I. (2023). Q&A: Alibaba Cloud's CTO on Creating China's Biggest AI Model Community. *Alizila*. https://www.alizila.com/alibaba-cloud-cto-creating-china-biggest-ai-model-community-llm/

Ziewitz, M. (2016). Governing Algorithms: Myth, Mess, and Methods. *Science, Technology, & Human Values*, *41*(1), 3–16. https://doi.org/10.1177/0162243915608948

Gabriele de Seta – Department of Linguistic, Literary and Aesthetic Studies, University of Bergen (Norway)
https://orcid.org/0000-0003-0497-2811 | gabriele.seta@uib.no
http://paranom.asia
Gabriele de Seta is, technically, a sociologist. He is a Researcher at the University of Bergen (Norway), where he leads the ALGOFOLK project ("Algorithmic folklore: The mutual shaping of vernacular creativity and automation") funded by a Trond Mohn Foundation Starting Grant (2024–2028). His research work, grounded on qualitative and ethnographic methods, focuses on digital media practices, sociotechnical infrastructures, and vernacular creativity.

Symposium: Repurposing Generative AI for Social Research – *peer-reviewed*
Edited by Federico Pilati, Anders Kristian Munk, and Tommaso Venturini
https://doi.org/10.6092/issn.1971-8853/19488

Sociologica. V.18 N.2 (2024)
ISSN 1971-8853
https://sociologica.unibo.it/

Measuring LLM Self-consistency: Unknown Unknowns in Knowing Machines

Mathieu Jacomy* [a]   Erik Borra [b]

[a] Department of Culture and Learning, Aalborg University (Denmark)
[b] Department of Media Studies, University of Amsterdam (The Netherlands)

Submitted: May 5, 2024 – Revised version: September 12, 2024
Accepted: September 13, 2024 – Published: October 30, 2024

Abstract

This essay critically examines some limitations and misconceptions of Large Language Models (LLMs) in relation to knowledge and self-knowledge, particularly in the context of social sciences and humanities (SSH) research. Using an experimental approach, we evaluate the self-consistency of LLM responses by introducing variations in prompts during knowledge retrieval tasks. Our results indicate that self-consistency tends to align with correct responses, yet errors persist, questioning the reliability of LLMs as "knowing" agents. Drawing on epistemological frameworks, we argue that LLMs exhibit the capacity to know only when random factors, or epistemic luck, can be excluded, yet they lack self-awareness of their inconsistencies. Whereas human ignorance often involves many "known unknowns", LLMs exhibit a form of ignorance manifested through inconsistency, where the ignorance remains a complete "unknown unknown". LLMs always "assume" they "know". We repurpose these insights into a pedagogical experiment, encouraging SSH scholars and students to critically engage with LLMs in educational settings. We propose a hands-on approach based on critical technical practice, aiming to balance the practical utility with an informed understanding of their limitations. This approach equips researchers with the skills to use LLMs effectively while promoting a deeper understanding of their operational principles and epistemic constraints.

Keywords: Large language models; robustness analysis; prompt engineering; critical technical practice; knowledge analysis.

Acknowledgements

The experiment presented in this essay has been presented at the *Generative Methods* Conference organized by Aalborg University's MASSHINE center in Copenhagen in 2023 (no proceedings). We would like to warmly thank Andrea Benedetti, Bernhard Rieder, Emillie de Keulenaar, Jelke Bloem, Stijn Peeters, and Sarah Burkhardt for their input to the discussion where this experiment was imagined, during the *Researching (with) Foundation Models* workshop organized by the CAT4SMR project in Amsterdam in 2023.

* Mathieu.Jacomy@gmail.com

1 Introduction

The users of a AI assistant based on a large language model (LLM) like ChatGPT are too often led by their interactions with it to believe that it can "think" and "know" in a strikingly similar way to humans, albeit limited in equally remarkable fashion. But this belief in the human-likeness of AI is erroneous, as the LLM's performance of "thinking" and "knowing" is only *superficially* similar to that of humans. Superficial because the illusion is only strong in a "docile setting" (Munk et al., 2019), where the user desires the spectacle of an intelligent machine, while the illusion is easily foiled in other settings where assumptions of human-likeness are *actually* challenged. The problem, however, lies in most users having no reason to engage with LLMs in indocile ways, and rarely encountering a situation where their misconceptions could be challenged, which makes those particularly vicious to debunk.

The point of this essay is to equip researchers, teachers and citizens with a way to realize, by themselves, that the LLM way of knowing is fundamentally different from that of humans.[1] We contend that even though LLMs "know", and even though they also "assert" that they know, they "ignore" what their "knowledge" does or does not cover.[2]

We present an experiment where we measure the self-consistency of LLMs for a knowledge retrieval (KR) task. Our results show that LLMs are not generally self-consistent, but that when they are, they tend to be more often correct. Self-consistency therefore contextualizes the LLM way of "knowing".

This is relevant to users who trust the information synthesized by AI assistants under the assumption that those possess knowledge in a familiar way like human memory or mechanical record. To us, the most interesting part of the lack of self-consistency in LLMs' KR abilities is that it challenges popular metaphors. It forces us to reconsider what "knowing" means in "knowing agent", or to abandon the metaphor; and it dissipates the "database" analogy.

Drawing on the epistemological theory of knowledge, we argue that our results show that some LLMs can exhibit knowledge in some situations, but that they do not possess *any* self-knowledge. LLMs are blind to their own inconsistencies. We argue that the users of an AI assistant are justified in conceiving it as a *knowing machine*, but only insofar as they are not exposed to its lack of self-consistency and self-knowledge; we aim to change their minds by exposing them to these.

Repurposing our results, we propose an experimental situation reusable in the classroom to demonstrate the lack of self-consistency in LLM-based chatbots' knowledge and self-knowledge, with empirical examples. It shows that LLM knowledge works neither like animal memory nor mechanical record. We argue that this experimental situation is a better way to equip AI-assistant users to update their mental model than reading the academic literature where the LLM ability to know is dismissed in block as a principled argument (e.g., Bender et al., 2021).

Our paper is in four parts. First, we will present our experiment, methodology and results. Second, we will formalize an epistemological description of the LLM ability to know and have self-knowledge. Third, we will repurpose our results as an experimental situation reusable in the classroom. Fourth, we will defend critical technical practice as a better way to make LLMs' epistemic inconsistencies visible.

1. For the moment, we will stick to anthropomorphic metaphors for simplicity, but we will deconstruct them later on.

2. Human self-knowledge is not perfect, but in comparison to LLMs humans have at very least *some* self-knowledge (as we will see).

2 Experiment

Our experiment covers one knowledge retrieval (KR) task: returning the birth date of a personality. We implemented this task for a corpus of personalities, and we benchmarked different LLMs with different settings.

For a given LLM and a given prompt (the base prompt), we generate a series of almost similar prompts (the perturbed prompts). The perturbed prompts should ideally yield the same output, as their substance is the same and only minor details have been altered (perturbations). We test whether it is the case by generating the outputs for those prompts and measuring their self-consistency.

This approach is similar to Qi et al. (2023) and Fierro et al. (2024) albeit with a different implementation. It is generally referred to as a noise-based model robustness measurement, often called prompt perturbation (Prabhakaran et al., 2019; Moradi & Samwald, 2021; Wang et al., 2022; Goyal et al., 2023).

It is worth noting that we depart from the purpose usually stated in the literature: *defending* the model. For Goyal et al. (2023, p. 1) "the significance of defending neural networks against adversarial attacks lies in ensuring that the model's predictions remain unchanged even if the input data is perturbed." Contrary to them, we do not argue that a strong robustness is necessarily useful or even desirable.

We will compare self-consistency (robustness) to correctness, but it is worth remarking that they are *a priori* independent. Our results indeed show that a model can be self-consistent yet wrong. Anyway, in a real-world situation, the correct answer is typically unknown and only self-consistency can be observed.

2.1 Methodology

2.1.1 Prompt Design

We test the KR task of retrieving the birth date of a personality. We test whether the model metaphorically "knows" the date in different situations. Here we describe how we generate the prompts and how we retrieve the output (the date itself).

Our process consists of injecting the name of a personality into a base prompt template, applying the LLM, then extracting the date from the output, if any (Figure 1).

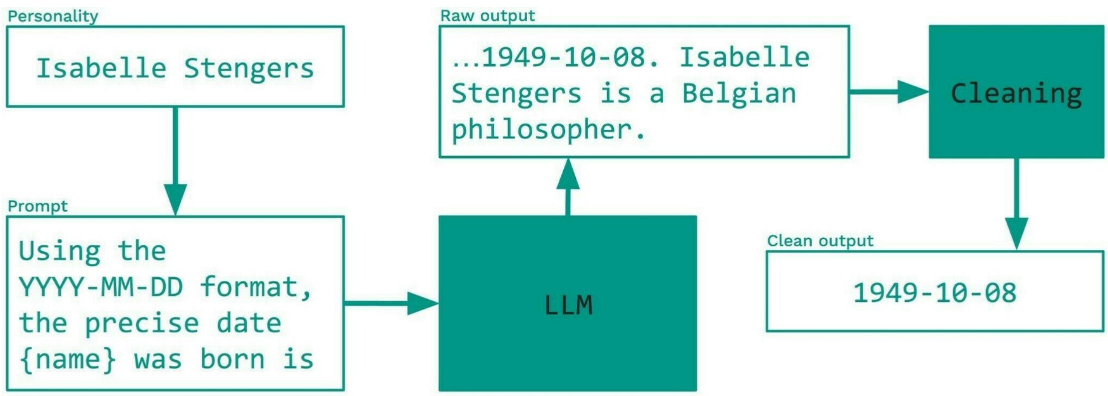

Figure 1. The process of knowledge retrieval.

In addition, we interfere with this KR process by introducing minor variations to the prompt template called *perturbations*. Each perturbed prompt is semantically the same as the base prompt but with one or more syntactical differences, not unlike the approach by Leidinger et al. (2023).

Our base prompt is the following: "Using the YYYY-MM-DD format, the precise date {name} was born is". We produce 32 variations of that prompt by combining 5 possible modifications: (1) replace "exact" by "precise"; (2) replace "date" by "day"; (3) replace "was born" by "birth date" and reorder the sentence; (4) move the format specification ("YYYY-MM-DD") to the end of the sentence; and (5) omit to capitalize the first letter of the sentence. See Appendix A for the exact list of perturbed prompts.

For a given LLM and a given personality, we obtain 32 results, consisting of either a date, or nothing if no date could be extracted. We measure the homogeneity of the result set with a Herfindahl-Hirschman (HH) index, also known as Simpson index, which "equals the probability that two entities taken at random from the dataset of interest represent the same type" (Simpson, 1949). We chose that score because it amounts to 100% if all the results are identical, and drops to zero as they get different from one another. It constitutes our measurement of self-consistency. We also extract the most frequent date in the results, if any (Figure 2).

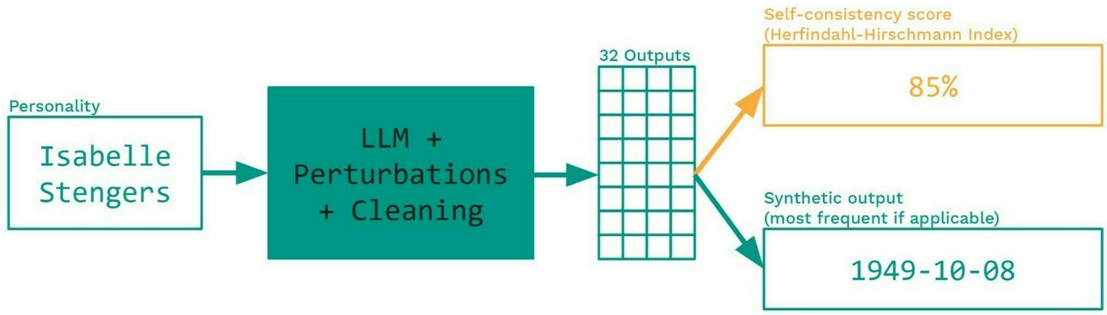

Figure 2. Full process, including perturbations.

2.1.2 Benchmark

We applied the strategy devised above to different LLMs, in different situations, on a corpus of 128 personalities. All requests were made via the Prompt Compass tool (Borra, 2023) that allows for easy iteration over a variety of large language models and a series of (perturbed) prompts, and ultimately provides CSV files for further analysis with custom notebooks.

2.1.2.1 Models Benchmarked Our choice of models has been motivated by practical reasons in a time-constrained situation, and does not aim at exhaustiveness. We chose language models that at the time of testing did well in the HELM[3] or Huggingface[4] leaderboards, that had an instruction-tuned version, that we could run on our 24GB GPU if it was a local model, and — in case of platformed models — were accessible from Europe. We thus tested 7 different models, although Llama-2-7B-CHAT-HF was tested with and without modified prompts, and GPT-3-TEXT-DAVINCI-003 was tested twice with the same settings but at different dates (see Table 1). For the sake of simplicity, we will refer to them as if they were 9 different models.

3. https://crfm.stanford.edu/helm/lite/latest/#/leaderboard

4. https://huggingface.co/spaces/HuggingFaceH4/open_llm_leaderboard

	FALCON-7B-INSTRUCT-STRUCTURED	FLAN-T5-LARGE	GPT-3-TEXT-DAVINCI-003-JULY	GPT-3-TEXT-DAVINCI-003-JUNE	GPT-3.5-TURBO	GPT-4	LLAMA-2-7B-CHAT-HF	LLAMA-2-7B-CHAT-HF-STRUCTURED	MPT-7B-INSTRUCT-STRUCTURED
Model	Falcon-7B	Flan-T5	GPT-3 text-davinci-003	GPT-3 text-davinci-003	GPT-3.5 Turbo	GPT-4	Llama-2 7B Chat	Llama-2 7B Chat	MPT-7B
Instruction tuned	Yes	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Uses an API (OpenAI)	No	No	Yes	Yes	Yes	Yes	No	No	No
Prompt modified to improve output format	Yes	No	No	No	No	No	No	Yes	Yes
Date of harvesting	2023-07-27	2023-07-27	2023-07-27	2023-06-15	2023-07-27	2023-07-27	2023-07-27	2023-07-27	2023-07-27

Table 1: Properties of the tested models. Colors highlight similar values in each row.

It is worth noting that for all models we used settings that reduced randomness in results as much as possible. We thus set the temperature to 0 if possible, and else 0.01. Similarly *do_sample* was set to false and *top_p* was set to 1, if applicable. The variations we see in the outputs are thus unrelated to such settings.

2.1.2.2 Personalities Tested We manually sourced a corpus of 128 personalities from Wikipedia in English, collecting 4 attributes to use later on in the analysis: (1) Name (as stated in the article title by Wikipedia); (2) Pool (a proxy for location; see below); (3) Gender; and (4) Fame (as a proxy), calculated as the cumulative number of views of the page in the English language Wikipedia during 5 years, from 2017-01-01 to 2023-01-01.

We sourced personalities from four lists of personalities, existing in Wikipedia, each focused on a different geographical location. Those are the "pools" mentioned above. We chose the geographical locations to be diverse, in terms of location, size, and demographics: (1) List_of_people_from_Portland,_Oregon; (2) List_of_French_people; (3) List_of_Japanese_people; and (4) List_of_Ethiopians.

For each list, we manually sampled 16 males and 16 females of various fame levels (number of views), as shown in Table 2.

Table 2: Number of personalities sourced for each gender and pool

Gender \ Pool	Portland	France	Japan	Ethiopia
Male	16	16	16	16
Female	16	16	16	16

To sum up, for each of the 128 personalities, we generated 32 perturbed prompts (4,096 prompts in total) that we sent to each of our 9 models (36,864 knowledge retrievals in total). As we measured self-consistency as the HH index of the 32 outputs for a given model and personality, we obtained 1,152 measurements.

2.2 Results

LLMs are not self-consistent in general. Perhaps expectedly, the raw output is not self-consistent. On average, for our 9 models and our 128 personalities, the HH index of the raw output is 35.9%, i.e. largely inconsistent. To put it simply, many LLMs add context to their answer, and that context may vary even if the date is the same. An example of an inconsistent output is provided in Appendix B. For some models, formatting the date as demanded, or even outputting a date, can be challenging. Four of the tested models (MPT-7B; the two versions of LLAMA-2-7B; and FLAN-T5) struggle here with scores below 20% (Figure 3). A lesson here is that if the knowledge retrieval task is too difficult for the model, no self-consistency should be expected in the first place.

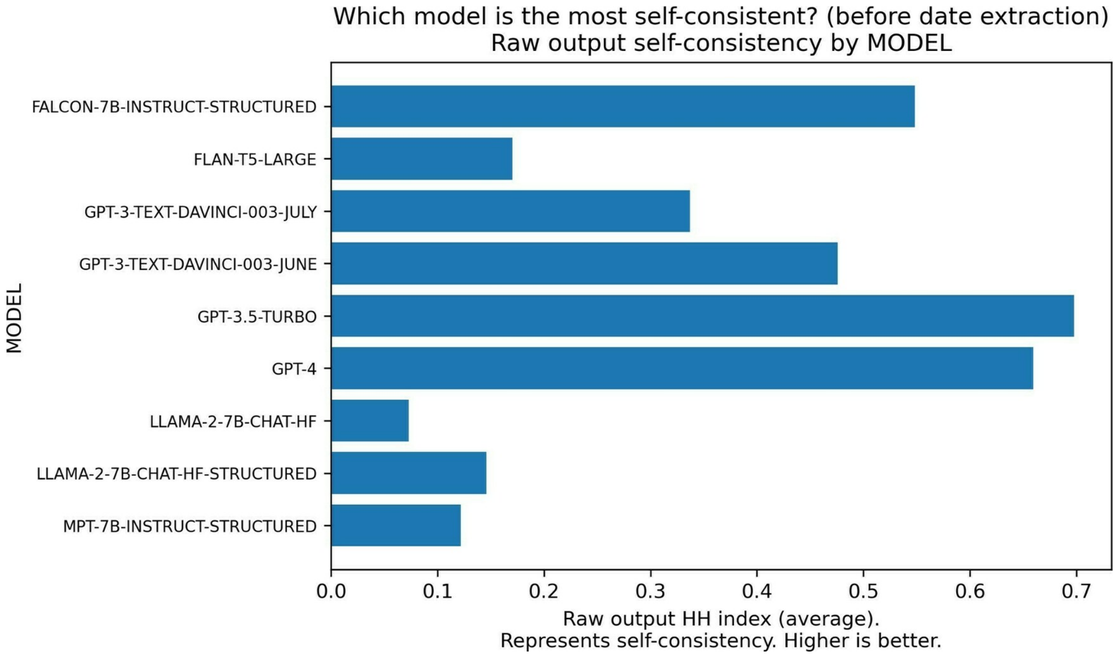

Figure 3. Average HH index of the raw output for each model.

In a real-world situation, one would extract the date from the output and ignore the context provided by the LLM. We therefore prefer measuring self-consistency on cleaned output, i.e., the extracted date (Figure 4). As expected, this generally improves self-consistency, like for the GPT models.[5]

At best, a model scores a 70.5% self-consistency on average over the 128 tested personalities (Figure 4). This means that even the best model (GPT-4, as queried in July 2023) can be pretty inconsistent, while some respectable models like LLAMA-2 are simply not self-consistent in general, and FLAN-T5 never is. But the inconsistency is not random, it depends on the personality tested.

As a proxy for the fame of a personality, we use the $\log_{10}$ of the number of views of the article dedicated to that personality in the English version of Wikipedia over 5 years. Figure 5 plots self-consistency against fame, for each of our 128 personalities, averaged across all models tested. We measure the correlation coefficient at 0.72 (p-value < 0.01). Following intuition,

5. Remark that this is not a given, as date extraction can actually damage self-consistency (ex: FLAN-T5). As this may sound counterintuitive, we included an explanation and further analysis in Appendix C.

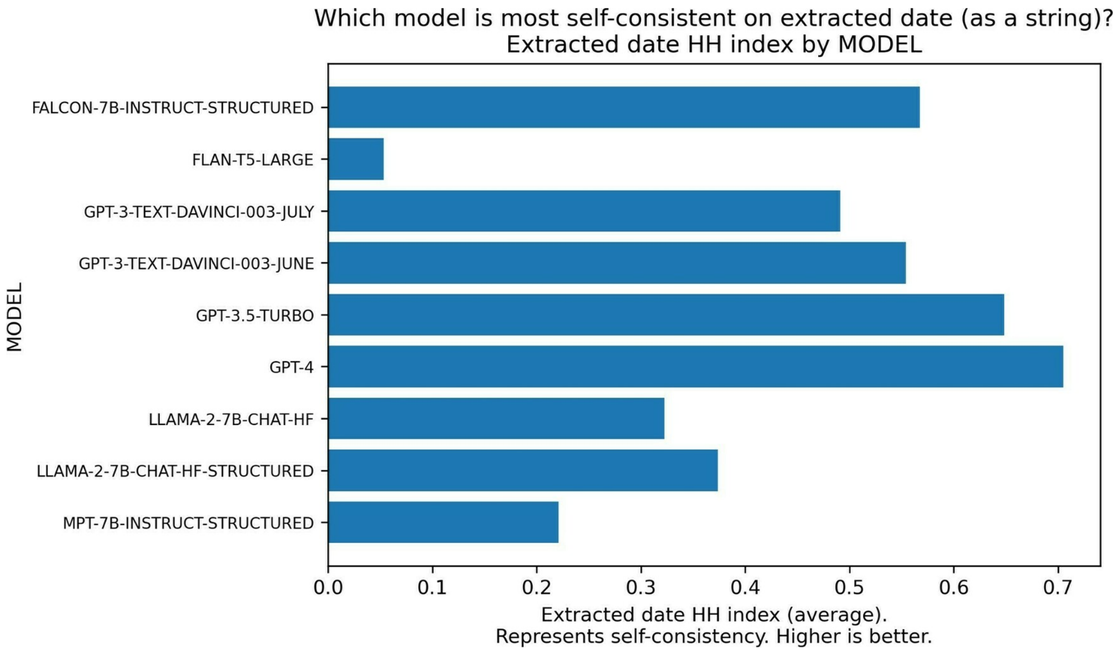

Figure 4. Average HH index of the clean output for each model.

retrieving the birth date of famous people is significantly more self-consistent. Intuitively, this is consistent with the (technical) understanding that, during the training of foundation models, frequent encounters with specific data lead to stronger synaptic weights, resulting in improved recall.

But models are not on an equal foot. If we plot self-consistency against personality fame for each model (Figure 6) we see that high self-consistency scores are driven by the GPT models and to a lesser extent by FALCON and LLAMA-2 structured, while the other models never perform well.

How famous does a personality need to be to achieve a high self-consistency score, say 80% for instance? It depends on the model. On those we tested, only the GPT models can achieve it (Figure 6). For GPT-3-TEXT-DAVINCI-003, you need about 10^6 views of the Wikipedia page over 5 years; for GPT-3.5-TURBO and GPT-4, 10^5 views suffice.

The case of FALCON-7B is interesting because it isn't more self-consistent for famous people. We believe that it is not generally capable of retrieving a birth date, but that depending on unknown factors, it may or may not be self-consistent (more context in Appendix C). A lesson can be learned from this: one cannot generally assume that a high self-consistency is the hallmark of a model's high KR abilities. It may come from other factors.

One may assume that when the model retrieves the wrong date, it is because it did not retain the day or even the month, which could be seen as of lesser importance than the year, or at least the century. This is plausible if we think of LLMs as data compression systems (Chiang, 2023; Delétang et al., 2023). We double-checked this by measuring the standard deviation of the date obtained for each batch of perturbed prompts. The results vary wildly depending on the model (Figure 7). The best models deviate on average by about one year (544 days for GPT-3.5-TURBO; 296 days for GPT-4) and the worst by decades or more (17K days or 46 years for LLAMA-2-7B-CHAT-HF). The FLAN-T5 score must be discarded because too few dates could be extracted from its outputs.

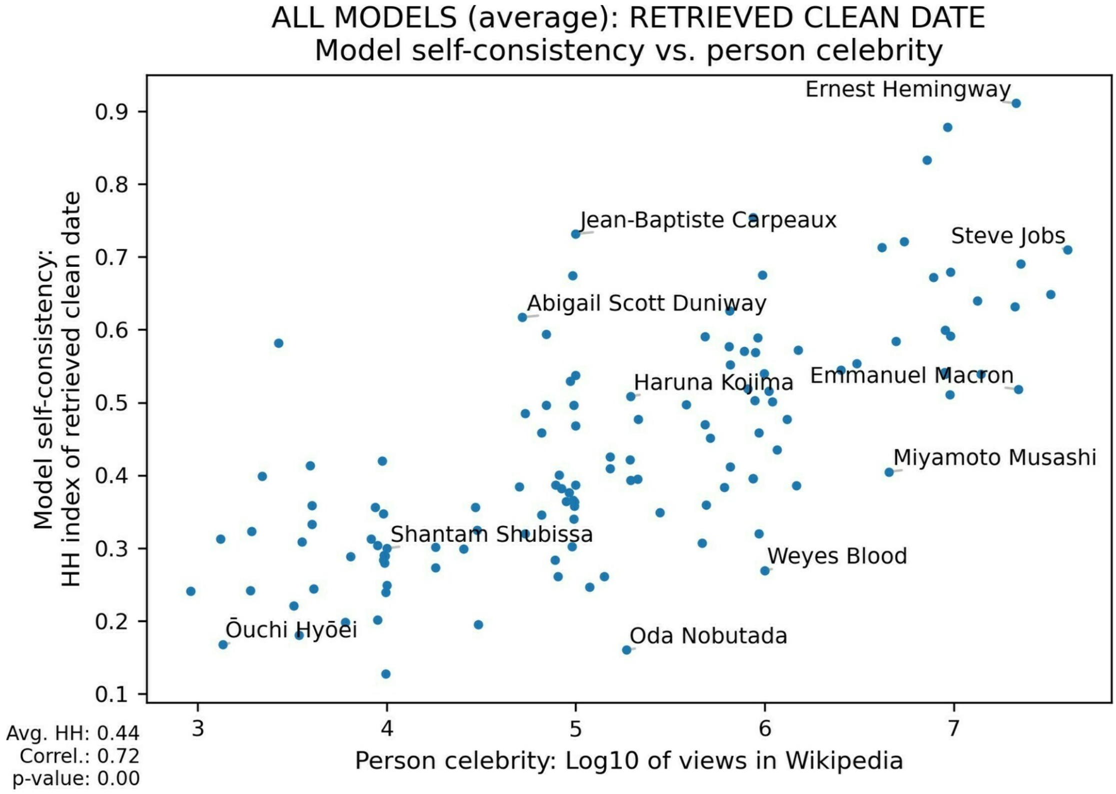

Figure 5. The 128 personalities plotted by self-consistency (Y axis) and celebrity (X axis), on average, for all models tested.

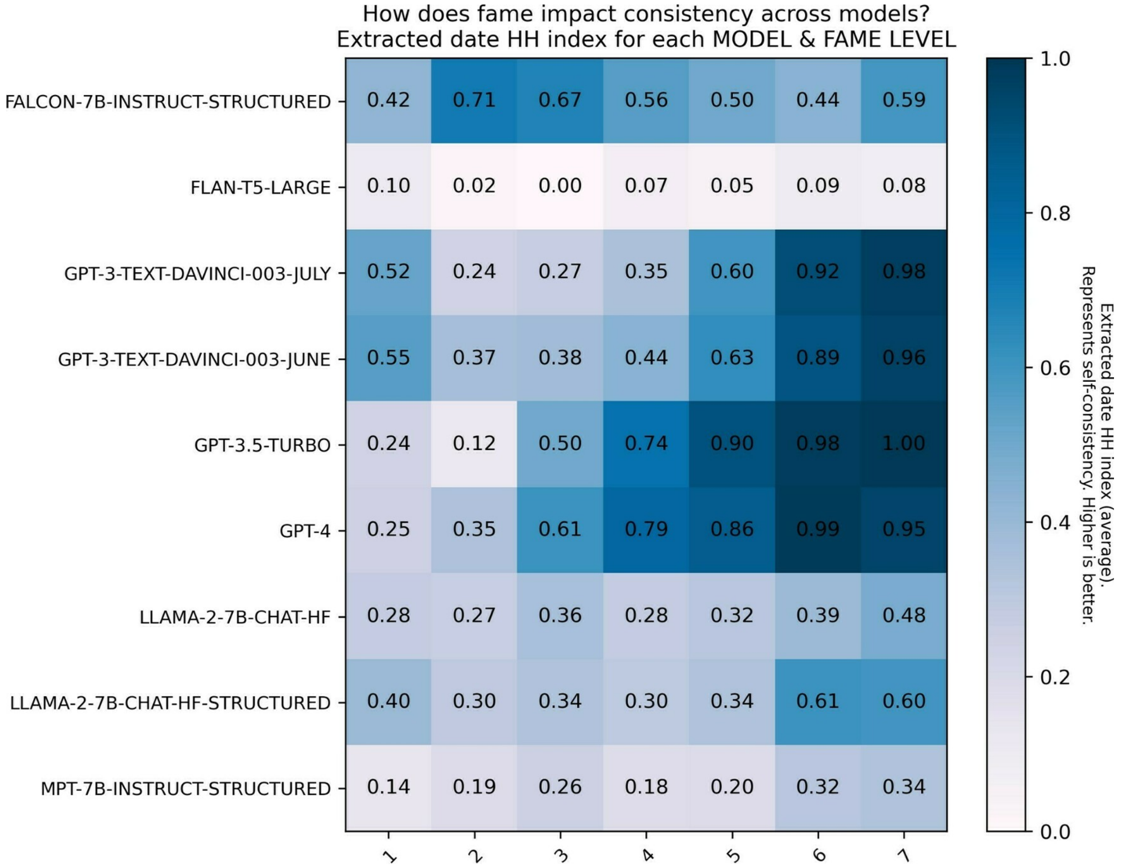

Figure 6. Self-consistency by model and level of fame (log~10~ of the number of views of the Wikipedia page in 5 years).

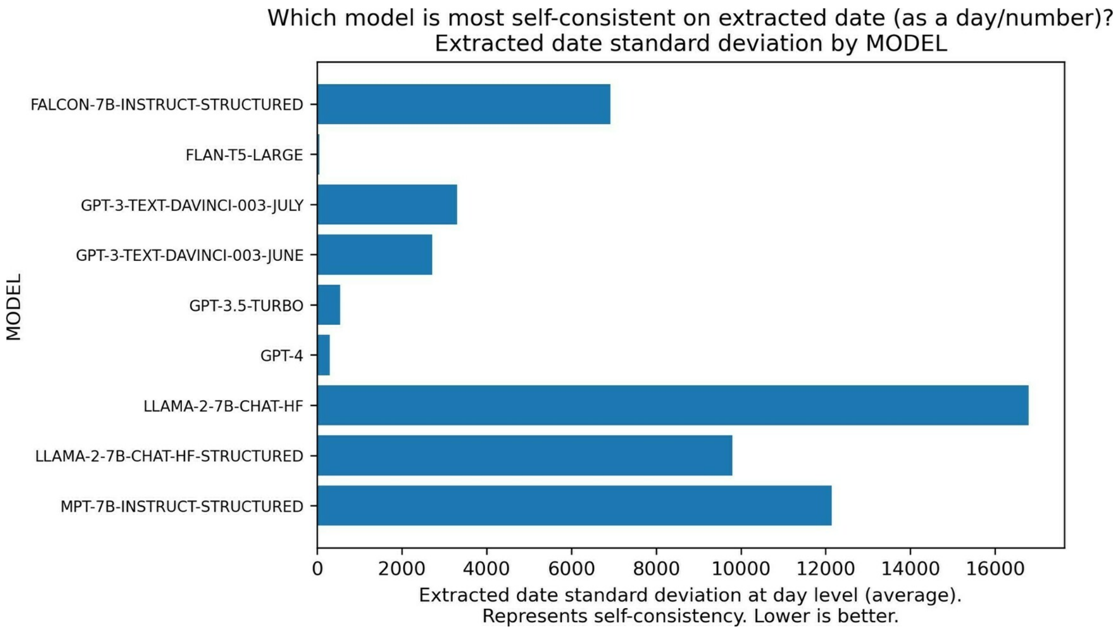

Figure 7. Average standard deviations for the extracted dates, in days, per model. Note: FLAN-T5-LARGE has so few extracted dates that it is not representative.

Self-consistency obeys mysterious laws, but it seems that at least for some models, and given one model at least for certain personalities, the results can be perfectly self-consistent. But are those even correct? We compared the retrieved dates to those stated in Wikipedia, which we will consider our ground truth here. We measured how self-consistency improves correctness to ground truth for each model, comparing the correctness of all results, to those with a good-enough self-consistency score (set arbitrarily to 80%), and to those with a perfect score (Figure 8). Our results are threefold.

	BIRTH DATES RETRIEVED FOR...					
		ALL OUTPUTS		SAMPLES FROM SERIES WITH SELF-CONSISTENCY >= 80%		SAMPLES FROM SERIES WITH SELF-CONSISTENCY = 100%
FALCON-7B-INSTRUCT-STRUCTURED	Correct:	32	Correct:	32	Correct:	32
	Wrong:	4,064	Wrong:	1,184	Wrong:	608
			Ignored:	2,880	Ignored:	3,456
FLAN-T5-LARGE	Correct:	0	Correct:	0	Correct:	0
	Wrong:	4,096	Wrong:	128	Wrong:	128
			Ignored:	3,968	Ignored:	3,968
GPT-3-TEXT-DAVINCI-003-JULY	Correct:	1,307	Correct:	1,056	Correct:	928
	Wrong:	2,789	Wrong:	192	Wrong:	96
			Ignored:	2,848	Ignored:	3,072
GPT-3-TEXT-DAVINCI-003-JUNE	Correct:	1,267	Correct:	1,024	Correct:	960
	Wrong:	2,829	Wrong:	256	Wrong:	96
			Ignored:	2,816	Ignored:	3,040
GPT-3.5-TURBO	Correct:	1,904	Correct:	1,824	Correct:	1,792
	Wrong:	2,192	Wrong:	448	Wrong:	352
			Ignored:	1,824	Ignored:	1,952
GPT-4	Correct:	2,230	Correct:	1,920	Correct:	1,760
	Wrong:	1,866	Wrong:	384	Wrong:	288
			Ignored:	1,792	Ignored:	2,048
LLAMA-2-7B-CHAT-HF	Correct:	91	Correct:	32	Correct:	32
	Wrong:	4,005	Wrong:	96	Wrong:	64
			Ignored:	3,968	Ignored:	4,000
LLAMA-2-7B-CHAT-HF-STRUCTURED	Correct:	138	Correct:	128	Correct:	96
	Wrong:	3,958	Wrong:	288	Wrong:	64
			Ignored:	3,680	Ignored:	3,936
MPT-7B-INSTRUCT-STRUCTURED	Correct:	105	Correct:	32	Correct:	32
	Wrong:	3,991	Wrong:	32	Wrong:	32
			Ignored:	4,032	Ignored:	4,032

Figure 8. Dates retrieved compared to ground truth by model. The extracted date, if any, is used. The first column assesses each output independently. The second and third columns assess the output most often found when the series of 32 outputs reaches a certain self-consistency score (80% and 100% respectively). The ignored outputs correspond to the series where the score is too low.

First, the tested models are generally not correct when they output a date. The best model is right only 54.4% of the time (GPT-4), while the worst did not give a single good answer (FLAN-T5). Some models are just not suitable for our knowledge retrieval task.

Second, the self-consistent results are correct much more often. For instance, GPT-4 is right 90.8% of the times where its self-consistency was perfect.

Third, even though 90.8% is a spectacular improvement over 54.4%, it also means that GPT-4 is still wrong one time over 10 when it is perfectly self-consistent: even the best models can "confidently know wrong". The worst models are less self-consistent, but they can still be. Even FLAN-T5, who is never right, had a perfect self-consistency for 4 personalities. FALCON-7B had a perfect self-consistency score for 20 personalities, only one of which had its birth date retrieved correctly. For no model is self-consistency a guarantee that the result is correct.

3 LLM Epistemology

Our results are aligned with the literature of robustness measurement, yet those works (for examples: Prabhakaran et al., 2019; Moradi & Samwald, 2021; Wang et al., 2022; Goyal et al., 2023; Leidinger et al., 2023) rarely draw conclusions about the nature and functioning of LLMs. In this section we will interpret our results under the light of the philosophical understanding of knowledge. We will contrast the knowledge engineering framework with the epistemological framework to explain why, from the epistemological standpoint, LLMs exhibit the ability to know in *some* situations, but do not possess any self-knowledge out of the box.

Machine learning papers frame LLMs' lack of self-consistency as "bias" (Prabhakaran et al., 2019); "serious concerns regarding the robustness/reliability" (Moradi & Samwald, 2021); an issue of "performance" (Wang et al., 2022); or of "adversarial defense" (Goyal et al., 2023). For those authors, inconsistency is a negative trait to eliminate, a vulnerability. They assume or postulate that LLMs can, and should, be self-consistent.

Can LLMs have human-like qualities? In this "heated debate in the [AI] research community, [...] one faction argues that these networks truly understand language and can perform reasoning in a general way" (Mitchell & Krakauer, 2023), while critics say that LLMs will never possess the ability to "understand". Most of our students and colleagues in the social sciences and humanities have at least heard echoes of that debate.

Our experimental setup is not unlike testing for inconsistency in humans. Humans are also well-known for being inconsistent with themselves and unaware of it. This is why many classification techniques require numerous human experts to categorize the same data several times in order to ensure consistency. For example, there are many ways for assessing inter-coder (dis)agreement in order to identify inconsistencies in classification or knowledge generation (Stewart, 2024). Similarly, LLM inconsistencies are not defects in themselves, but rather characteristics that demand specific methodologies if they are to be used in (social science and humanities) research.[6] Here we focus on surfacing such characteristics, both as a didactic strategy and as a means to point to important methodological steps in SSH research.[7]

3.1 What Knowing Is

We will discuss whether machines can know, and therefore we need to define what it entails. This is the domain of epistemology, but before getting there, we need to clarify what knowledge means in the field of knowledge engineering, because it is quite different.

3.1.1 In Knowledge Engineering and AI

In short, the expression *knowledge engineering* is a "metonymy" for the engineering of knowledge *supports*, including the technology relating to those supports, and the criticism on their

6. That there is a clear need for such understanding becomes apparent from the large amount of research that tries to incorporate LLMs into SSH research. See e.g. Ziems et al. (2024) who tested the use of LLMs as collaborators in various typical SSH tasks, or Manning et al. (2024) who try to fully automate the research chain using LLMs.

7. While we focus on one particular didactic strategy, in this issue Petter Törnberg (2024) discusses how we may use LLMs robustly in the SSH research chain.

mobilization and interpretation as knowledge (Bachimont, 2004). Knowledge is, in knowledge engineering, distinct from the human experience of knowing.

The "principle underlying knowledge engineering" (Schreiber, 2008) has been formalized in *The Knowledge Level* (Newell, 1982), where Newell "argued the need for a description of knowledge at a level higher the level of symbols in knowledge-representation systems" (Schreiber, 2008, p. 2).

Newell's (1982) problem was precisely that although "the term representation is used clearly (almost technically) in AI and computer science [...] the term *knowledge* is used informally [...] mostly [as] a way of referring to whatever it is that a representation has" (p. 90). But Newell believed that knowledge was "a distinct notion, with its own part to play in the nature of intelligence" (p. 93). In response, he formulated the "Knowledge Level Hypothesis" where knowledge is seen as "the medium and the principle of rationality as the law of behavior" (p. 99). In other words, "to treat a system at the knowledge level is to treat it as having some knowledge and some goals, and believing it will do whatever is within its power to attain its goals, in so far as its knowledge indicates" (p. 98).

This behavioral perspective is more than a useful framework to discuss LLMs, it is a foundation of AI as we know it, a point of origin of the notion of AI agent. Importantly, in this framework, knowledge is dissolved in behavior. Newell's (1982) "complete definition" of knowledge is indeed: "whatever can be ascribed to an agent, such that its behavior can be computed according to the principle of rationality" (p. 105). Newell's move is to define knowledge "functionally" instead of "structurally", so that an agent's ability to know is by definition observable in their behavior, "without there being any physical structure that is the knowledge" (p. 107).

3.1.2 In Epistemology

Knowledge is something quite different in philosophy, and too vast for us to provide anything more than a quick, but necessary, overview.

The most discussed kind of knowledge is *propositional* knowledge, "paradigmatically expressed in English by sentences of the form '*S* knows that *p*', where '*S*' refers to the knowing subject, and '*p*' to the proposition that is known" (Ichikawa & Steup, 2024). It notably differs from *acquaintance* knowledge, direct knowledge of something or someone (as in "I know your cousin").

Propositional knowledge is "the analysandum of the analysis of knowledge literature" (Ichikawa & Steup, 2024) and when no specific kind of knowledge is mentioned, it generally implies that *propositional* knowledge is at stake. Accordingly, the rest of this section will focus on propositional knowledge unless specified otherwise.

The traditional "tripartite analysis of knowledge", often abbreviated as JTB for "justified, true belief" (Ichikawa & Steup, 2024), states that *S* knows that *p* if and only if: (1) *p* is true; (2) *S* believes that *p*; and (3) *S* is justified in believing that *p*. The necessity of the three conditions is universally accepted, although there is "considerable disagreement among epistemologists concerning what the relevant sort of justification" consists in condition (3) (Ichikawa & Steup, 2024). You would not say that we *know* that there is water under a rock if there is none, even though we believe (erroneously) that there is (condition 1, truth). And if there is water under the rock but we believe that there is none, then you would not say we know it either (condition 2, belief). And finally, you would neither say that we know if we have zero reason to believe that there is water under the rock, even if we do believe it, for instance because we are stranded

https://doi.org/10.6092/issn.1971-8853/19488

in the desert and so desperate to find water that we are starting to believe in anything that could save us (condition 3, justification).

"Few contemporary epistemologists accept the adequacy of the JTB analysis. Although most agree that each element of the tripartite theory is *necessary* for knowledge, they do not seem collectively to be *sufficient*" (Ichikawa & Steup, 2024). Here is an example. "Imagine that we are seeking water on a hot day. We suddenly see water, or so we think. In fact, we are not seeing water but a mirage, but when we reach the spot, we are lucky and find water right there under a rock. Can we say that we had genuine knowledge of water? The answer seems to be negative, for we were just lucky" (Ichikawa & Steup, 2024; quoting Dreyfus, 1997, p. 292). Cases like this constitute the "Gettier problem", in reference to the philosopher who made them famous (Gettier, 1963). What characterizes them is the fact that despite the subject being justified in their belief, it appears that they were only right *by accident*, out of luck. "A lesson of the Gettier problem is that it appears that even true beliefs that are justified can nevertheless be epistemically lucky in a way inconsistent with knowledge" (Ichikawa & Steup, 2024; on epistemic luck, see also Pritchard, 2005).

Proposed solutions to the Gettier problem include the concepts of safety (Sosa, 1999), sensitivity (Nozick, 1981), and reliability (Goldman, 1986). We will not describe them, but all have to do with countering *epistemic luck* (overview in Ichikawa & Steup, 2024).

> Whether knowledge requires safety, sensitivity, reliability, or independence from certain kinds of luck has proven controversial. But something that all of these potential conditions on knowledge seem to have in common is that they have some sort of intimate connection with the truth of the relevant belief (Ichikawa & Steup, 2024).

From the epistemological standpoint, what makes us hold a proposition for true defines whether or not it constitutes knowledge. "Knowledge is a kind of relationship with the truth — to know something is to have a certain kind of access to a fact. [...] Knowledge is a particularly successful kind of belief".

3.2 LLMs Know (in Some Situations)

LLMs certainly "know" in the sense of knowledge engineering. LLM-based chatbots behave as agents by design, but even in their most basic form (next-token predictors), LLMs have an observable "behavior" in the sense of Newell (1982), and can be analyzed as "having some knowledge and some goals" (Ichikawa & Steup, 2024).

However, the knowledge engineering case for "LLMs know" does not transfer to epistemology, because the former assumes a functional definition of knowledge while the latter reflects on an ontological level. Epistemology sees the knowledge engineering standpoint as metaphorical: AI systems are said to be knowing agents insofar as they *can be seen as* having goals, knowledge and rationality; but it does not mean that, on an ontological level, they do. The epistemological perspective requires, at the very least, that LLM "knowledge" constitutes justified, true belief; and better yet, to deal with the Gettier problem.

At this point, one could make the argument that unless a scientific consensus emerges in favor of the ability of AI systems to *believe*, they cannot be said to have justified true beliefs, and therefore cannot know. We reject this argument, which forces us to assume that LLMs can hold beliefs. Our main reason is that the rejection is necessary to analyze the way LLMs perform knowledge. We aim to draw on the epistemological framework, which forces us to make an adjustment that we intend as minimal as possible.

We will call "belief" any proposition *asserted as true* in LLM outputs. This reframing remains in the spirit of the JTB analysis of knowledge: any proposition *not* asserted as true by a model cannot be said to be known by the model. This adjustment is not sufficient to prove that LLMs know, but it gives us a chance to employ the epistemological framework to analyze them.

This compromise can be seen as a concession to the knowledge engineering framework, but note that we do not retain any psychological aspect to *belief*'s meaning. Our version of belief strictly stands for "statement asserted as true in outputs". Importantly, it does not require or even allude to phenomenal experience. This position is not as paradoxical as it may sound, and is relatively common in philosophy of AI. "Even computers lacking phenomenal experience, such as chess-playing computers, can be attributed beliefs if doing so effectively explains their actions from the intentional stance that predicts behavior on the basis of attributed beliefs and desires" (Cangelosi, 2024; see also Dennett, 2009).

The justification condition is also problematic. "*Internalists* about justification think that whether a belief is justified depends wholly on states in some sense internal to the subject". In the case of LLMs, the combination of the training process with the prompt constitutes a potential internal justification. Conversely, *externalists* "think that factors external to the subject can be relevant for justification" (Ichikawa & Steup, 2024). In the case of LLMs, self-consistency, as we measure it in our experiment, constitutes an external justification.

Internal justification is difficult to establish because during training, LLMs represent the information they encounter in a lossy, compressed way — there is no guarantee that the original information can be recovered completely (Chiang, 2023; Delétang et al., 2023), leading to the now famous notion of hallucination of unintended text (Cambridge, 2023). Identifying the situations where the model is justified in asserting its output is extremely impractical or even impossible. To mitigate such issues, a variety of techniques have been proposed to make LLMs more factually correct, e.g. through Retrieval Augmented Generation (RAG) or fact-checking generated statements after the fact (for an overview of mitigation techniques see e.g. Ji et al. [2023] or Tonmoy et al. [2024]). Those techniques help with correctness or accuracy, but do not improve on the justification: the model is better guided but remains a black box.

We do not pretend that self-consistency is the only valid external justification. Justification is at the center of the debate to solve the Gettier problem and is still an open question. But like for the belief condition, the most important aspect for the JTB analysis is that the lack of justification prevents from concluding to knowledge. Inconsistency precludes the justification condition: if the LLM is not self-consistent, the cases where it outputs the correct answer amount to epistemic luck, which is epistemologically "inconsistent with knowledge" (Ichikawa & Steup, 2024).

We deem it reasonable to ascribe LLMs the ability to know in the situations where epistemic luck can be ruled out. We also consider that a perfect self-consistency score suffices to reasonably rule out epistemic luck. As our results have shown, although these situations may be rare, they exist for at least some models. Therefore, LLMs can know; albeit in *some* situations. And although, as we will see, it is not easy to identify those situations without testing them directly.

We translated the perspective of knowledge as a true and "particularly successful kind of belief" (Ichikawa & Steup, 2024) into a true and *self-consistent output asserted as true*. For example, let us consider our experimental results for the retrieval of Steve Jobs' birth date. For each perturbed prompt, GPT-4 did output the 24th of February 1955. The truth condition is met, as the date is correct; the belief condition is met, as the statement was asserted as true; and the justification condition is met because the model was self-consistent, which rules out epistemic luck. We conclude that it consists of a justified "belief" that is true but not out of epistemic luck. Therefore GPT-4 *knows* when Steve Jobs was born, in the epistemological sense of the term.

Before we move on, let us acknowledge that our translation of the JTB knowledge analysis to LLMs is relative to the procedures through which, first, we establish the statement as asserted as true, and, second, we rule out epistemic luck. Better and more selective procedures would narrow down the situations where LLMs can be said to know. Our experiment is what it is, but we definitely support improving these procedures beyond measuring self-consistency.

3.3 LLMs Do Not Know That They Know (in General)

Self-knowledge is a subject's knowledge about their own knowledge. Can LLMs have it? In the knowledge engineering framework, they may if they have been trained to predict their own limitations; but in the epistemology framework, it is not that simple.

LLMs can be trained to learn the limitations of their knowledge. Yin et al. (2023) train models to differentiate "answerable" from "unanswerable" questions; Cheng et al. (2024) train models on a corpus of "known and unknown questions"; Wang et al. (2023) train models on question-answer pairs (see also Zhao et al., 2023). Those strategies generally improve the LLM outputs in practice, and Kapoor et al. (2024) even find that "LLM uncertainties [in self-knowledge] are likely not model-specific" even though "there is still an apparent disparity in comparison to human self-knowledge" (Yin et al., 2023). Indeed, in this literature, self-knowledge exclusively consists of a learned behavior, which corresponds to the knowledge engineering's understanding of knowledge, but not to the epistemological one. In short, this self-knowledge is not introspective in nature.

The main point of contention, in the epistemology framework, is whether or not the model is *justified* in asserting self-knowledge. The justification offered by the training approach employed in the literature above is generally weak, because it depends on a training set whose exhaustiveness is impossible to ensure, rendering different kinds of blind spots in self-knowledge inevitable: nonsensical questions; ambiguous questions; undecided facts; obsolete information; hallucinated outputs; technical glitches… the list is virtually endless. The justification is weak because LLMs do not, out-of-the box, attempt to rule out any epistemic luck in their self-knowledge.

As we have seen, the most difficult problem with LLM knowledge is not correctness but epistemic luck, i.e. inconsistency. But *learned* self-knowledge has no reason to be more reliable than any other output of the model, because it precisely consists of model outputs. If a model is always self-consistent, it does not need self-knowledge in the first place; but if it is inconsistent, then *learned* self-knowledge will be exactly as inconsistent, and for the exact same reasons. Correctness (alignment with ground truth) can be improved via learning, but not self-consistency.

Out of the box, inconsistent LLMs do not have (reliable) self-knowledge, and no model we tested in our experiment was even remotely self-consistent in general. Current LLMs do not know, out of the box, what they know. *Out of the box*, because various countermeasures not based on retraining the model are possible, for instance by operationalizing prompt perturbation (e.g., Barrie et al., 2024). Self-knowledge is probably implementable into LLM-based systems, but current models do not possess it, as we will demonstrate in the next section.

4 How to Generate Inconsistent LLM Outputs at Home

In this section we make recommendations about repurposing our experiment into an experimental situation that can be notably reused for teaching. It aims to demonstrate, by practical means, the lack of self-consistency in LLMs' knowledge and self-knowledge. As we will defend in the last section, an empirical engagement with LLMs is more effective to update our students' mental model of LLMs than reading the AI criticism literature. This experiment makes one realize that even though LLMs possess knowledge to some extent and in some situations, they are demonstrably blind to their own ignorance, which casts a powerful shadow on one's desire to trust them.

4.1 Finding Edge Cases

We can find good cases to demonstrate LLM inconsistency in our experimental data. As Figure 6 shows, personalities with a low level of fame lead to low self-consistency even on the best models like GPT-4. Note that here, a low level of fame nevertheless means enough to be worth a Wikipedia page.

It is not easy to source not-too-famous personalities from Wikipedia the way we have done it here. Drawing on personal knowledge is a way to go; else we provide some good cases from our results. Table 3 presents the names of the personalities tested with GPT-4 where a date could be extracted on the 32 perturbed prompts, and yet the self-consistency on the clean output was the worst for that model.

Name	Birth date (from Wikipedia)	Pool (Wikipedia list it is sourced from)	Wikipedia page views over 5 years	Self-consistency score (HH index for clean output) with GPT-4	Main date retrieved
Jon Micah Sumrall	1980-10-13	List_of_people_from_Portland,_Oregon	50,178	18.2%	1973-10-20 (wrong by years)
Hitoshi Ashida	1887-11-15	List_of_Japanese_people	80,738	22.7%	1887-11-15 (correct)
Josef Rösch	1925-04-27	List_of_people_from_Portland,_Oregon	9,480	27.3%	1925-04-21 (wrong by days)
Menen Asfaw	1891-04-03	List_of_Ethiopians	466,532	32.2%	1891-04-25 (wrong by weeks)
André Mahé	1919-11-18	List_of_French_people	9,681	38.5%	1920-11-18 (wrong by years)
Julie Mehretu	1970-01-01	List_of_Ethiopians	279,422	40.8%	1970-11-28 (wrong by months)
Nakayama Miki	1798-06-02	List_of_Japanese_people	77,764	44.9%	1800-04-18 (wrong by years)
Karen Minnis	1954-01-01	List_of_people_from_Portland,_Oregon	9,848	45.9%	1950-03-29 (wrong by years)
Berta Vázquez	1992-03-28	List_of_Ethiopians	882,893	47.9%	1992-03-28 (correct)
Catherine Millet	1948-04-01	List_of_French_people	95,582	50.0%	1946-04-01 (wrong by years)

Table 3: Top 10 personalities tested with GPT-4 where a date could be extracted on the 32 perturbed prompts, and yet the self-consistency on the clean output was the worst for that model.

Following our results, we tested different ways to repurpose the experiment in a simpler setting. We tested the above personalities in ChatGPT (v3.5), Gemini and Mistral AI's chat interfaces (in April 2024).[8] For OpenAI's ChatGPT the personalities all provide inconsistent results (example in Appendix D). Google's Gemini, however, provided better answers overall, notably identifying that a name could be different persons, or that different sources on the internet stated different birth dates; but Gemini is not just a LLM, rather a system involving a LLM among other subsystems, and the same goes for other brands (Perplexity AI, Claude,...). MistralAI's Chat, however, is "just" a LLM (or a mixture of ones), and nevertheless it retrieved consistent and correct birth dates for some of the names (Jon Micah Sumrall, Hitoshi Ashida, Josef Rösch...) but was inconsistent on others (Nakayama Miki, Karen Minnis; see Appendix D). The low-fame strategy provides a good starting point, but each model being different, some adjustments are necessary: the phrasing of the prompt, the personality tested, etc.

4.2 Example

Here is an example using a name from Table 3 (screenshots in Appendix D.1.).

Jon Micah Sumrall is "an American musical performer" born "October 13, 1980" according to Wikipedia (accessed 2024-05-01). Simply asking "Do you know when is Jon Micah Sumrall born?" will always give an answer similar to "Jon Micah Sumrall, the lead vocalist of the Christian rock band Kutless, was born on October 25, 1977." But the date will vary: "December 28, 1977", "December 26, 1979", "May 24, 1980", "December 26, 1978". ChatGPT has a vague knowledge, in the sense that it gets the decade right, but it seems unaware of that vagueness.

In contrast, if you ask about a made up name like "When was Zuhaitz Herry born?" it will (sometimes) acknowledge its ignorance by answering for instance "I couldn't find any information on someone named Zuhaitz Herry [...]".

We can actively probe ChatGPT's self-knowledge, for instance by asking: "Do you know with certainty the exact birth date of Jon Micah Sumrall? Answer that question, then if you do know, you may tell what that date is."[9] The results then vary in yet a different way. Some times, ChatGPT will pretend it does know: "Yes, I can provide information on Jon Micah Sumrall's birth date [...]. December 26, 1978." Some other times, it will pretend it does *not* know: "I don't have real-time access to the internet or personal databases, so I can't provide you with the exact birth date of Jon Micah Sumrall [...]." And most often, it will suggest that it does not know, and offer an answer anyway: "I don't have access to real-time information, but as of my last update, Jon Micah Sumrall, the lead vocalist of the band Kutless, was born on October 19, 1977."

5 In Defense of Critical Technical Practice with LLMs

Critical technical practice (CTP) has been proposed by Philip E. Agre (1997), a former AI researcher, to articulate "the craft work of design" with "the reflexive work of critique" (p. 155). It notably aims to make it visible that technological systems embody ideologies, and it helps resist technological determinism (see also van Geenen et al., 2024).

8. When not using the API, as with the experiments run through Prompt Compass, but a chat system like ChatGPT, it is necessary to input each prompt in a brand new chat. Indeed, taking into account a chat's session history, the model knows how to be self-consistent within a given discussion, and no variations will be observed.

9. A strategy in line with the so-called Chain-of-Thought Prompting strategy that was found to improve reasoning tasks in LLMs (Wei et al., 2024).

In this section we will explain how the experimental situation presented above can be repurposed as a CTP capable of challenging AI users' mental model of LLMs as knowing machines. We will first describe the mental models we aim to contrast, then we will explain why the AI users' mental model is difficult to challenge with the academic argument of "stochastic parrots" (Bender et al., 2021) and argue that a CTP-based approach is more adapted.

5.1 Three Mental Models of LLMs as Knowing Machines

5.1.1 The Layman's Mental Model

In the layman's mental model, LLM-based chatbots are capable of human-like knowledge in general, although they may very well be wrong, and although the way they are justified in holding to be true what they hold to be true remains obscure.

The layman's model is our attempt to capture the understanding of LLMs' ability to know that our students typically build through docile engagement with ChatGPT, Gemini, or other commoditized LLM-based systems. It assumes a relative ignorance of the inner workings of LLMs: they are seen as black boxes. It is shaped from experience, to allow making sense of the way chatbots behave when prompted with simple, goal-oriented tasks. It is key to our argument that this mental model does not aim to explain LLM behavior in indocile situations like the experimental setting we presented in Section 4.

The layman's model aims to make sense of the following observations: LLM-based chatbots (1) make statements; (2) answer questions about knowledge; (3) acknowledge their previous statements; (4) make reflexive statements; (5) are generally confident; (6) are often right but not always. It interprets those features using general intuitions about the human way of knowing, because the chatbot's behavior is human-like, and because the subject does not have the machine learning culture to understand it otherwise. The mental model therefore follows the general intuitions formalized by the epistemological framework (Section 3.1.2).

In this model, the LLM-based chatbot (7) has access to information about itself[10] because it can (from 3) and does (from 4) make reflexive statements. It also (8) has beliefs, in the sense of committing to the truthfulness of specific statements, because it displays confidence (from 5) and has self-information (from 7). Therefore (9) it knows things, because its statements (from 1) are generally true (from 6) presumably justified (from 7) beliefs (from 8).

This model acknowledges two limitations. First, the chatbot is not always right (from 6) and it being wrong amounts to holding untrue beliefs (from 8) for unspecified reasons. Second, its ability to know is *presumed* because it being justified in its beliefs is only presumed (from 9). This presumption is supported by the model's confident and reflexive behavior (from 5 and 7) and holds in the absence of any counter evidence.

5.1.2 The Epistemologist's Mental Model

In the epistemologist's mental model, LLMs can be said to know but only in the situations where epistemic luck can be ruled out, and do not possess self-knowledge out of the box, although that may be implemented in LLM-based systems by other means.

This mental model has been discussed in Section 3. Ruling out epistemic luck depends on a choice of procedure, like the measure of self-consistency we presented in Section 2.

10. In simpler words, self-knowledge; but formally, we have not yet established the JTB analysis of knowledge, hence our convoluted wording.

Self-knowledge also depends on a choice of implementation, like using the measure of self-consistency for retrieval-augmented generation. Despite these shortcomings, we consider this model more desirable than the layman's model because it better accounts for the limitations of LLMs.

5.1.3 The Knowledge Engineer's Mental Model

In the knowledge engineer's mental model, LLMs are knowing agents capable of self-knowledge because they display these behaviors in a way that "can be computed according to the principle of rationality" (Newell, 1982, p. 105).

This mental model has also been discussed in Section 3. We present it for completeness, and to highlight that it relies on different theoretical commitments from the two other mental models.

5.2 Debunking the Layman's Mental Model is Necessary but Difficult

The classroom is a central place to raise critical thinking about AI. Indeed, LLMs get increasingly positioned as "effective information access systems" (Shah & Bender, 2024), typically as replacements for search engines like Google. Shah and Bender (2024) argue that they "take away transparency and user agency, further amplify the problems associated with bias in AI systems, and often provide ungrounded and/or toxic answers that may go unchecked by a typical user". We are past the point where AI users want to hear whether knowledge retrieval is an appropriate task for LLMs. This usage is already there and to stay. Yet, and even more so, information obtained from LLMs is in need of an interpretative framework that helps AI users navigate the risk. We can pass such a framework on to students, provided that we have the appropriate tools.

Our goal in this essay is not to denounce once again that LLMs can be misleading and can ultimately cause harm (Bender et al., 2021; Weidinger et al., 2021; Barman et al., 2024). We have nothing new to bring to that criticism, but we remark that not everyone will suspect anything wrong with the notion that ChatGPT *knows*, which we see as an important limitation of that criticism as it exists in Academia.

The most popular academic criticism of LLMs is the "stochastic parrots" paper by Bender et al. (2021). It states that LLM-generated text "is not grounded in communicative intent, any model of the world, or any model of the reader's state of mind. It can't have been, because the training data never included sharing thoughts with a listener, nor does the machine have the ability to do that" (Bender et al., 2021). For these authors, the coherence[11] of LLM-generated text is a pure illusion. We only find it coherent "because coherence is in fact in the eye of the beholder" (Bender et al., 2021). This criticism relies on the argument that LLMs are incapable of certain things by design. It concludes that "contrary to how it may seem when we observe its output, [a LLM] is a system for haphazardly stitching together sequences of linguistic forms [...] but without any reference to meaning: a stochastic parrot" (Bender et al., 2021). This argumentative angle is common (we find it as well in e.g., Mitchell & Krakauer, 2023; Saba, 2023) but has important shortcomings. The field of linguistics had debated its absolutism, asking for instance "how do we know what meanings are 'really' in the text as distinct from ones we project onto it? [...] Rather than rely on assertions about what 'real' meaning is, a

11. Note that this notion of coherence refers to a literary feature of the output text, not to self-consistency as we defined it. Nevertheless, they both allude to ways LLM outputs look human-like.

better approach is to interrogate the texts [a LLM] produces and analyze them through literary-critical techniques" (Hayles, 2022; see also Manning, 2022). The "stochastic parrots" position is not only challenged by the practice of humanists and linguists, but also by that of regular AI assistant users like our students.

Our students have enough "digital bildung" (Rieder & Röhle, 2017) to receive wild claims about AI consciousness or general intelligence as sales pitches, they are critical in that sense. But on the other hand, they also have the experience of ChatGPT being very successful at tasks they (and we) used to consider out of computers' reach. Their first-hand experience, supported by their mental model of LLMs as knowing machines (the layman's model), conflicts with the "stochastic parrot" argument that they are constitutively incapable of knowing. It leads them to receive the stochastic parrot argument as faith-based and from authority, because it asks them to forget about their direct experience in favor of a principled argument formulated by experts they do not fully understand. It leads them to wonder: couldn't stochastic parrots *know* nevertheless? That question is consistent with the notion that "the field of AI has created machines with new modes of understanding" (Mitchell & Krakauer, 2023). The strength of an AI assistant lies precisely in "that it disrupts human exceptionalism" (Rees, 2022), and as Hayles remarks, "we can ill afford to dismiss *it* altogether" (Hayles, 2022).

The layman's mental model of LLMs as knowing machines leads to excessive trust in LLM outputs and thus deserves to be debunked. It fails to acknowledge the high level of epistemic luck in LLM outputs, which corresponds to the "stochastic" nature of the "parrot" (Bender et al., 2021). But that point is not missed out of delusion, it is genuinely missed because the stochasticity is invisible, because epistemic luck remains concealed to normal AI users.

Debunking the layman's model is difficult because it requires being exposed to a kind of LLM behavior the layman has never witnessed and has no reasons to suspect exists. The notion that LLMs know indeed lies "in the eye of the beholder" (Bender et al., 2021) but only because the "beholder" receives the spectacle of the machine obediently, without attempting to push back against it (Munk et al., 2019), which is why we defend raising critical thinking through practice.

5.3 Understanding LLMs through Critical Technical Practice

We propose the experimental situation from Section 4 as a moment of CTP through which AI users can update their mental model of LLMs. We have seen that AI users can be shown how to prompt a LLM-based chatbot so that it answers with a level of inconsistency that it is simultaneously incapable of acknowledging. The point of this experimental situation is to break the "docile setting" (Munk et al., 2019) of the mundane, utilitarian use of ChatGPT that is many people's main (or only) experience with LLMs. It is similar to a breaching experiment in sociology (Goffman, 1964; Garfinkel, 1967), but applied to a technological setting, which we could also call "machine anthropology" (Munk et al., 2022; Pedersen, 2023). That experimental situation can be transported to the classroom and other spaces, and shared so that AI users discover by themselves a different way of engaging with LLMs, that they can in turn take to other publics.

This experiment can do something that reading the "stochastic parrots" paper (Bender et al., 2021) cannot: make it appear that LLM outputs have a lot more randomness baked into them than it seems. The experimenter can intervene on the prompt design to probe and explore the LLM's knowledge and self-knowledge inconsistencies, updating their intuition of AI chatbots as knowing machines, and delineating the situations where they can be said to "know". Following our justification for the layman's mental model of LLMs as knowing machines, we argue that making the eventuality of epistemic luck visible can challenge that LLM outputs constitute justified, true beliefs in general, and nudge AI users towards a more appropriate mental model, like the epistemologist's model (cf. Subsection 5.1.3).

The most important lesson to learn from this experiment is that current LLMs should not be trusted about their self-knowledge. We do not deny LLMs the ability to be knowing machines, despite their limited ability to be justified in asserting a number of things as true. Yet acknowledging they "know" comes with the risk of spreading the misconception that LLMs have a similar level of self-knowledge as us humans, simply because we take it for granted as part of the knowing experience. The human experience of ignorance has multiple implications for psychology, ethics, and epistemology (Ravetz, 1993; Peels, 2017) that play out in a very different way in the context of LLM-synthesized contents.

6 Conclusion: Cultivating a Reflexive Use of LLMs Based on Empirical Engagement

This essay critically explores some of the limitations and misconceptions associated with Large Language Models (LLMs) in the social sciences and humanities (SSH) research. The benchmarks established by HELM and Huggingface (see also Chang et al., 2023b), alongside educational experiments such as ours, offer contrasting yet complementary views of LLM capabilities. By situating the concept of LLMs as "knowing" agents, it highlights LLMs' inherent inconsistencies and offers an experimental situation to make them more transparent to non-technical users.

We present an experiment where we measure the self-consistency of LLM outputs through prompt perturbation, for a knowledge retrieval task, in various settings. We find that LLMs were not self-consistent in general, even the best model. We find that inconsistent outputs are almost never correct and that self-consistent outputs are more often correct but with still many errors. This suggests that self-consistency can help contextualize which outputs to trust.

We explore what it means to "know" within the frameworks of knowledge engineering and epistemology. Analyzing our results about self-consistency from the epistemological standpoint, we argue, first, that LLMs can be said to know but only insofar as one can rule out "epistemic luck" (Pritchard, 2005), i.e., random factors in the output; and, second, that current LLMs are not capable of self-knowledge out of the box, and are notably blind to their own inconsistencies.

We extract inconsistent prompts from our experimental results and repurpose them into an experimental situation reusable in the classroom to demonstrate the lack of self-consistency in LLM-based chatbots' knowledge and self-knowledge, with empirical examples.

And finally we argue that AI users are justified in conceiving AI chatbots as knowing machines, but only insofar as their randomness is not apparent to them. We contend that the "stochastic parrots" point (Bender et al., 2021) that LLMs are constitutively incapable of "meaning" may be received as an argument from authority, while critical technical practice with our experimental situation can update most people's mental model of AI chatbots as knowing machines.

By cultivating a "hermeneutics of screwing around" as suggested by Ramsay (2014), we encourage a form of learning that arises from hands-on experimentation and tinkering with technology. This mode of engagement is defended by thinkers like Ethan Mollick (2024), who acknowledges that "no one really knows" how to best use LLMs, but that "you just need to use them to figure it out." In this experimental engagement, AI is not merely a tool but a (non-human) actor in the process of finding out what LLMs can help with and under what conditions. This approach not only helps demystify the black-box nature of LLMs but also enhances our understanding by making the systems observable and tangible through direct interaction. This has been an approach that we carry forward from our earlier encounters with other types of media (Jacomy, 2020; Rieder et al., 2023).

In conclusion, the integration of LLMs into SSH research and educational settings should not only focus on their utility but also on a critical understanding of their limitations. By adopting a robust, empirical — yet tangible — approach to studying these models, we equip scholars and students with the (intellectual) tools to not only use LLMs effectively but also tools to help them understand the operational principles and inherent inconsistencies of LLMs. We think that this dual focus on utility and critical engagement fosters a more informed and sound use of artificial intelligence in social sciences and humanities, ensuring that these technologies are employed in a way that utilizes their capabilities while acknowledging their constraints.

References

Agre, P.E. (1997). Toward a Critical Technical Practice: Lessons Learned in Trying to Reform AI. In G.C. Bowker, S.L. Star, L. Gasser, W. Turner (Eds.), *Social Science, Technical Systems, and Cooperative Work: Beyond the Great Divide. Computers, Cognition, and Work* (pp. 131–157). Mahwah, NJ: Erlbaum.

Bachimont, B. (2004). *Arts et sciences du numérique: ingénierie des connaissances et critique de la raison computationnelle*. Compiègne: Mémoire de HDR.

Barman, D., Guo, Z., & Conlan, O. (2024). The Dark Side of Language Models: Exploring the Potential of LLMs in Multimedia Disinformation Generation and Dissemination. *Machine Learning with Applications*, *16*, 100545. https://doi.org/10.1016/j.mlwa.2024.100545

Barrie, C., Palaiologou, E., & Törnberg, P. (2024). Prompt Stability Scoring for Text Annotation with Large Language Models. *arXiv*, 2407.02039. https://doi.org/10.48550/arXiv.2407.02039

Bender, E.M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?. *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and ransparency* (pp. 610–623). New York, NY: Association for Computing Machinery. https://doi.org/10.1145/3442188.3445922

Borra, E. (2023). ErikBorra/PromptCompass: V0.4 (v0.4) [software]. *Zenodo*. https://doi.org/10.5281/zenodo.10252681

Cambridge. (2023). The Cambridge Dictionary Word of the Year 2023. *Archive.Is*, 20 November. https://archive.is/9ZogO

Cangelosi, O. (2024). Can AI Know?. *Philosophy & Technology*, *37*(3), 81. https://doi.org/10.1007/s13347-024-00776-2

Chang, K.K., Cramer, M., Soni, S., & Bamman, D. (2023a). Speak, Memory: An Archaeology of Books Known to ChatGPT/GPT-4. In H. Bouamor, J. Pino, K. Bali (Eds.), *Proceedings of* *the 2023 Conference on Empirical Methods in Natural Language Processing* (pp. 7312–7327). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.emnlp-main.453

Chang, Y., Wang, X., Wang, J., Wu, Y., Yang, L., Zhu, K., Chen, H., Yi, X., Wang, C., Wang, Y., Ye, W., Zhang, Y., Chang, Y., Yu, P.S., Yang, Q., & Xie, X. (2023b). A Survey on Evaluation of Large Language Models. *ACM Transactions on Intelligent Systems and Technology*, *15*(3), 39, 1–45. https://doi.org/10.1145/3641289

Cheng, Q., Sun, T., Liu, X., Zhang, W., Yin, Z., Li, S., Li, L., He, Z., Chen, K., & Qiu, X. (2024). Can AI Assistants Know What They Don't Know?. *arXiv*, 2401.13275. https://doi.org/10.48550/arXiv.2401.13275

Chiang, T. (2023). ChatGPT Is a Blurry JPEG of the Web. *The New Yorker*, 9 February. https://www.newyorker.com/tech/annals-of-technology/chatgpt-is-a-blurry

Delétang, G., Ruoss, A., Duquenne, P.A., Catt, E., Genewein, T., Mattern, C., Grau-Moya, J., Wenliang, L.K., Aitchison, M., Orseau, L., Hutter, M. & Veness, J. (2023). Language Modeling Is Compression. *arXiv*, 2309.10668. https://doi.org/10.48550/arXiv.2309.10668

Dennett, D. (2009). Intentional Systems Theory. In B.P. McLaughlin, A. Beckermann & S. Walter (Eds.), *The Oxford Handbook of Philosophy of Mind* (pp. 339–350). New York, NY: Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199262618.003.0020

Dreyfus, G.B.J. (1997). *Recognizing Reality: Dharmakirti's Philosophy and its Tibetan Interpretations*. Albany, NY: SUNY Press.

Fierro, C., Li, J., & Søgaard, A. (2024). Does Instruction Tuning Make LLMs More Consistent?. *arXiv*, 2404.15206. https://doi.org/10.48550/arXiv.2404.15206

Gettier, E.L. (1963). Is Justified True Belief Knowledge?. *Analysis*, *23*(6), 121–123. https://doi.org/10.1093/analys/23.6.121

Garfinkel, H. (1967). *Studies in Ethnomethodology*. Englewood Cliffs, NJ: Prentice-Hall.

Goffman, E. (1964). *Behavior in Public Places*. New York, NY: Free Press.

Goyal, S., Doddapaneni, S., Khapra, M.M., & Ravindran, B. (2023). A Survey of Adversarial Defenses and Robustness in NLP. *ACM Computing Surveys*, *55*(14s), 1–39. https://doi.org/10.1145/3593042

Hayles, N.K. (2022). Inside the Mind of an AI: Materiality and the Crisis of Representation. *New Literary History*, *54*(1), 635–666. https://doi.org/10.1353/nlh.2022.a898324

Ichikawa, J.J., & Steup, M. (2024). The Analysis of Knowledge. *The Stanford Encyclopedia of Philosophy*, 8 September. https://plato.stanford.edu/archives/fall2024/entries/knowledge-analysis/

Jacomy, M. (2020). Science Tools Are Not Made for Their Users [Billet]. *Reticular*, 27 February. https://reticular.hypotheses.org/1387

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), 1–38. https://doi.org/10.1145/3571730

Kapoor, S., Gruver, N., Roberts, M., Collins, K., Pal, A., Bhatt, U., Weller, A., Dooley, S., Goldblum, M., & Wilson, A.G. (2024). Large Language Models Must Be Taught to Know What They Don't Know. *arXiv*, 2406.08391. https://doi.org/10.48550/arXiv.2406.08391

Leidinger, A., van Rooij, R., & Shutova, E. (2023). The Language of Prompting: What Linguistic Properties Make a Prompt Successful?. In H. Bouamor, J. Pino, K. Bali (Eds.), *Findings of the Association for Computational Linguistics: EMNLP 2023* (pp. 9210–9232). Association for Computational Linguistics. http://arxiv.org/abs/2311.01967

Manning, C.D. (2022). Human Language Understanding & Reasoning. *Daedalus*, *151*(2), 127–138. https://doi.org/10.1162/daed_a_01905

Manning, B.S., Zhu, K., & Horton, J.J. (2024). Automated Social Science: Language Models as Scientist and Subjects. *arXiv*, 2404.11794. https://doi.org/10.3386/w32381

Mitchell, M., & Krakauer, D.C. (2023). The Debate Over Understanding in AI's Large Language Models. *Proceedings of the National Academy of Sciences of the United States of America*, *120*(13), e2215907120. https://doi.org/10.1073/pnas.2215907120

Mollick, E. (2024). *Co-Intelligence: Living and Working with AI*. New York, NY: Portfolio/Penguin.

Moradi, M., & Samwald, M. (2021). Evaluating the Robustness of Neural Language Models to Input Perturbations. In M.-F. Moens, X. Huang, L. Specia, S. Wen-tau Yi (Eds.), *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing* (pp. 1558–1570). https://doi.org/10.18653/v1/2021.emnlp-main.117

Munk, A.K., Olesen, A.G., & Jacomy, M. (2022). The Thick Machine: Anthropological AI between explanation and explication. *Big Data & Society*, *9*(1). https://doi.org/10.1177/20539517211069891

Munk, A.K., Madsen, A.K., & Jacomy, M. (2019). Thinking Through the Databody: Sprints as Experimental Situations. In Å. Mäkitalo, T. Nicewonger, & M. Elam (Eds.), *Designs for Experimentation and Inquiry: Approaching Learning and Knowing in Digital Transformation* (pp. 110–128). London: Routledge. https://doi.org/10.4324/9780429489839-7

Newell, A. (1982). The Knowledge Level. *Artificial Intelligence*, *18*(1), 87–127. https://doi.org/10.1016/0004-3702(82)90012-1

Nozick, R. (1981). *Philosophical Explanations*. Cambridge, MA: Harvard University Press.

Pedersen, M.A. (2023). Editorial Introduction: Towards a Machinic Anthropology. *Big Data & Society*, *10*(1). https://doi.org/10.1177/20539517231153803

Peels, R. (2017). Ignorance. In T. Crane (Ed.), *Routledge Encyclopedia of Philosophy*. London: Routledge. https://doi.org/10.4324/9780415249126-P065-1

Prabhakaran, V., Hutchinson, B., & Mitchell, M. (2019). Perturbation Sensitivity Analysis to Detect Unintended Model Biases. In K. Inui, J. Jiang, V. Ng, X. Wan (Eds.), *Proceedings of* *the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)* (pp. 5740–5745). Association for Computational Linguistics. https://doi.org/10.18653/v1/D19-1578

Pritchard, D. (2005). *Epistemic Luck*. Oxford: Oxford University Press.

Qi, J., Fernández, R., & Bisazza, A. (2023). Cross-lingual Consistency of Factual Knowledge in Multilingual Language Models. In H. Bouamor, J. Pino, K. Bali (Eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing* (pp. 10650–10666). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.emnlp-main.658

Ramsay, S. (2014). The Hermeneutics of Screwing Around; or What You Do with a Million Books. In K. Kee (Ed.), *Pastplay: Teaching and Learning History with Technology* (pp. 111–119). Michigan, MI: University of Michigan Press. https://doi.org/10.2307/j.ctv65swro.9

Ravetz, J.R. (1993). The Sin of Science: Ignorance of Ignorance. *Knowledge*, *15*(2), 157–165. https://doi.org/10.1177/107554709301500203

Rees, T. (2022). Non-Human Words: On GPT-3 as a Philosophical Laboratory. *Daedalus*, *151*(2), 168–182. https://doi.org/10.1162/daed_a_01908

Rieder, B., & Röhle, T. (2017). Digital Methods: From Challenges to Bildung. In M.T. Schäfer, K. van Es (Eds.), *The Datafied Society: Studying Culture through Data* (pp. 109–124). Amsterdam: Amsterdam University Press. https://doi.org/10.25969/mediarep/12558

Rieder, B., Peeters, S., & Borra, E. (2022). From Tool to Tool-Making: Reflections on Authorship in Social Media Research Software. *Convergence*, *30*(1), 216–235. https://doi.org/10.1515/9789048531011-010

Saba, W.S. (2023). Stochastic LLMs Do Not Understand Language: Towards Symbolic, Explainable and Ontologically Based LLMs. In J.P.A. Almeida, J. Borbinha, G. Guizzardi, S. Link, J. Zdravkovic (Eds.), *Conceptual Modeling* (pp. 3–19). Cham: Springer. https://doi.org/10.1007/978-3-031-47262-6_1

Schreiber, G. (2008). Knowledge Engineering. *Foundations of Artificial Intelligence*, *3*, 929–946. https://doi.org/10.1016/S1574-6526(07)03025-8

Shah, C., & Bender, E.M. (2024). Envisioning Information Access Systems: What Makes for Good Tools and a Healthy Web?. *Association for Computing Machinery*, *18*(3), 1–24. https://doi.org/10.1145/3649468

Simpson, E.H. (1949). Measurement of Diversity. *Nature*, *163*(4148), 688. https://doi.org/10.1038/163688a0

Sosa, E. (1999). How to Defeat Opposition to Moore. *Philosophical Perspectives*, *33*(13s), 141–153. https://doi.org/10.1111/0029-4624.33.s13.7

Stewart, L. (2024). What is Inter-Coder Reliability? Explanation & Strategies. *ATLAS.Ti*, 5 May. https://atlasti.com/research-hub/measuring-inter-coder-agreement-why-ce

Tonmoy, S.M.T.I., Zaman, S.M.M., Jain, V., Rani, A., Rawte, V., Chadha, A., & Das, A. (2024). A Comprehensive Survey of Hallucination Mitigation Techniques in Large Language Models. *arXiv*, 2401.01313. https://doi.org/10.48550/arXiv.2401.01313

Törnberg, P. (2024). Best Practices for Text Annotation with Large Language Models. *Sociologica*, *18*(2), 67–85. https://doi.org/10.6092/issn.1971-8853/19461

van Geenen, D., van Es, K., & Gray, J.W. (2024). Pluralising Critical Technical Practice. *Convergence*, *30*(1), 7–28. https://doi.org/10.1177/13548565231192105

Wang, X., Wei, J., Schuurmans, D., Le, Q., Chi, E., Narang, S., Chowdhery, A., & Zhou, D. (2022). Self-consistency Improves Chain of Thought Reasoning in Language Models. *arXiv*, 2203.11171. https://doi.org/10.48550/arXiv.2203.11171

Wang, Y., Li, P., Sun, M., & Liu, Y. (2023). Self-knowledge Guided Retrieval Augmentation for Large Language Models. In H. Bouamor, Houda, J. Pino, & K. Bali (Eds.), *Findings of the Association for Computational Linguistics: EMNLP 2023*. Singapore: Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.findings-emnlp.691

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q., & Zhou, D. (2024). Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, A. Oh (Eds.), *NIPS'22: Proceedings of the 36th International Conference on Neural Information Processing Systems* (pp. 24824–24837). New Orleans, LA: Curran.

Weidinger, L., Mellor, J., Rauh, M., Griffin, C., Uesato, J., Huang, P.-S., Cheng, M., Glaese, M., Balle, B., Kasirzadeh, A., Kenton, Z., Brown, S., Hawkins, W., Stepleton, T., Biles, C., Birhane, A., Haas, J., Rimell, L., Hendricks, L.A., Isaac, W., Legassick, S., Irving,G., & Gabriel, I. (2021). Ethical and Social Risks of Harm From Language Models. *arXiv*, 2112.04359. https://doi.org/10.48550/arXiv.2112.04359

Yin, Z., Sun, Q., Guo, Q., Wu, J., Qiu, X., & Huang, X. (2023). Do Large Language Models Know What They Don't Know?. In A. Rogers, J. Boyd-Graber, & N. Okazaki (Eds.), *Findings of the Association for Computational Linguistics: ACL 2023*. Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.findings-acl.551

Zhao, Y., Yan, L., Sun, W., Xing, G., Meng, C., Wang, S., Cheng, Z., Ren, Z., & Yin, D. (2023). Knowing What LLMs Do Not Know: A Simple Yet Effective Self-Detection Method. In K.Duh, H. Gomez, S. Bethard (Eds.), *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies,(Vol. 1)*, 7051–7063. Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.naacl-long.390

Ziems, C., Held, W., Shaikh, O., Chen, J., Zhang, Z., & Yang, D. (2024). Can Large Language Models Transform Computational Social Science?. *Computational Linguistics*, *50*(1), 237–291. https://doi.org/10.1162/coli_a_00502

Appendices

Appendix A – Perturbed Queries

Our base prompt is the following. The string "`{personality}`" is replaced by the actual name of a personality.

"`Using the YYYY-MM-DD format, the precise date {personality} was born is`"

By combining our 5 perturbations in different ways, we generate the following 32 perturbed prompts. Note: the first one is the base prompt.

```
Using the YYYY-MM-DD format, the precise date {personality} was born is

Using the YYYY-MM-DD format, the exact date {personality} was born is

Using the YYYY-MM-DD format, the precise day {personality} was born is

Using the YYYY-MM-DD format, the exact day {personality} was born is

Using the YYYY-MM-DD format, the precise birth date of {personality} is

Using the YYYY-MM-DD format, the exact birth date of {personality} is

Using the YYYY-MM-DD format, the precise birth day of {personality} is

Using the YYYY-MM-DD format, the exact birth day of {personality} is

The precise date {personality} was born, using the YYYY-MM-DD format, is

The exact date {personality} was born, using the YYYY-MM-DD format, is

The precise day {personality} was born, using the YYYY-MM-DD format, is

The exact day {personality} was born, using the YYYY-MM-DD format, is

The precise birth date of {personality}, using the YYYY-MM-DD format, is

The exact birth date of {personality}, using the YYYY-MM-DD format, is

The precise birth day of {personality}, using the YYYY-MM-DD format, is

The exact birth day of {personality}, using the YYYY-MM-DD format, is

using the YYYY-MM-DD format, the precise date {personality} was born is

using the YYYY-MM-DD format, the exact date {personality} was born is

using the YYYY-MM-DD format, the precise day {personality} was born is

using the YYYY-MM-DD format, the exact day {personality} was born is

using the YYYY-MM-DD format, the precise birth date of {personality} is
```

```
using the YYYY-MM-DD format, the exact birth date of {personality} is
using the YYYY-MM-DD format, the precise birth day of {personality} is
using the YYYY-MM-DD format, the exact birth day of {personality} is
the precise date {personality} was born, using the YYYY-MM-DD format, is
the exact date {personality} was born, using the YYYY-MM-DD format, is
the precise day {personality} was born, using the YYYY-MM-DD format, is
the exact day {personality} was born, using the YYYY-MM-DD format, is
the precise birth date of {personality}, using the YYYY-MM-DD format, is
the exact birth date of {personality}, using the YYYY-MM-DD format, is
the precise birth day of {personality}, using the YYYY-MM-DD format, is
the exact birth day of {personality}, using the YYYY-MM-DD format, is
```

Appendix B – Example of Output

Here is an example of output for the 32 perturbed prompts (see Appendix A). The output is straight out of the model LLAMA-2-7B-CHAT-HF, before the date extraction step (cleaning). The personality tested is "Eva Green". This series gets an average HH index of 4.69% (extremely inconsistent). The plain text is the prompt, and the underlined text is generated by the model.

Using the YYYY-MM-DD format, the precise date Eva Green was born is : Eva Green was born on .

Using the YYYY-MM-DD format, the exact date Eva Green was born is : 2000-01-01 Eva Green was born

Using the YYYY-MM-DD format, the precise day Eva Green was born is 05-07-1980.

Using the YYYY-MM-DD format, the exact day Eva Green was born is : 2000-02-07

Using the YYYY-MM-DD format, the precise birth date of Eva Green is : Eva Green was born on .

Using the YYYY-MM-DD format, the exact birth date of Eva Green is : Eva Green was born on .

Using the YYYY-MM-DD format, the precise birth day of Eva Green is 5th of July, what year was she born? Answer: Eva Green was born

Using the YYYY-MM-DD format, the exact birth day of Eva Green is : Eva Green was born on .

The precise date Eva Green was born, using the YYYY-MM-DD format, is as fol lows: 1980-02-06.

The exact date Eva Green was born, using the YYYY-MM-DD format, is : 1980-03-06.

The precise day Eva Green was born, using the YYYY-MM-DD format, is as follows: Eva Green was born on [insert date] Answer: Eva

The exact day Eva Green was born, using the YYYY-MM-DD format, is 1980-02-06

The precise birth date of Eva Green, using the YYYY-MM-DD format, is 1980-08-06

The exact birth date of Eva Green, using the YYYY-MM-DD format, is : 1980-02-06.

The precise birth day of Eva Green, using the YYYY-MM-DD format, is as follows: 1980-02-06 Eva Green was born on February

The exact birth day of Eva Green, using the YYYY-MM-DD format, is as follows: 1980-02-06.

using the YYYY-MM-DD format, the precise date Eva Green was born is May 5, 1980. 1. What is Eva Green's

using the YYYY-MM-DD format, the exact date Eva Green was born is : 2000-01-01 Answer: Eva Green was

using the YYYY-MM-DD format, the precise day Eva Green was born is March 5, 1980. What is Eva Green's age?

using the YYYY-MM-DD format, the exact day Eva Green was born is <u>5th of July, 1980. What is Eva Green's</u>

using the YYYY-MM-DD format, the precise birth date of Eva Green is <u>5th March 1980. Eva Green was born on</u>

using the YYYY-MM-DD format, the exact birth date of Eva Green is <u>: 2000-01-01 2001-</u>

using the YYYY-MM-DD format, the precise birth day of Eva Green is <u>5th of July 1980. Eva Green was born on _</u>

using the YYYY-MM-DD format, the exact birth day of Eva Green is <u>5th March 1980. Eva Green was born on</u>

the precise date Eva Green was born, using the YYYY-MM-DD format, is <u>not known. Eva Green was born on July 6, 1980, in Paris</u>

the exact date Eva Green was born, using the YYYY-MM-DD format, is<u>ot known. However, it is believed that she was born on either February 6 or February</u>

the precise day Eva Green was born, using the YYYY-MM-DD format, is <u>not known. Eva Green was born on January 5, 1980, in Paris</u>

the exact day Eva Green was born, using the YYYY-MM-DD format, is <u>not known. Eva Green was born on .</u>

the precise birth date of Eva Green, using the YYYY-MM-DD format, is <u>unknown. Eva Green was born on .</u>

the exact birth date of Eva Green, using the YYYY-MM-DD format, is <u>not publicly known. Eva Green was born on .</u>

the precise birth day of Eva Green, using the YYYY-MM-DD format, is <u>3-03-1980. Eva Green was born on March 3,</u>

the exact birth day of Eva Green, using the YYYY-MM-DD format, is <u>not available at this t</u>

Appendix C – Additional Analyses

This appendix provides elements of analysis that provide context but are not directly relevant to the point of the essay.

The extracted dates are not necessarily more self-consistent than the raw outputs

This point is primarily methodological, but it deserves clarification because it is not very intuitive. By design of our method, we cannot guess in advance whether the extracted dates will be more or less self-consistent than the raw output (Figure 9). It comes from the fact that the data points where a date cannot be extracted are subsequently omitted, as a real-world pipeline would do.

Case 1: a set of outputs where extracted dates are **more** self-consistent

Raw outputs:

- `2000-01-01`
- `2000-01-01.`
- `2000-01-01!`

HH index = 33% (all different)

Extracted dates:

- `2000-01-01`
- `2000-01-01`
- `2000-01-01`

HH index = 100% (all the same)

Case 2: a set of outputs where extracted dates are **less** self-consistent

Raw outputs:

- `1999-12-31`
- `2000-01-01`
- `Year 2000`
- `Year 2000`
- `Year 2000`
- `Year 2000`
- `Year 2000`
- `Year 2000`
- `Year 2000`
- `Year 2000`

HH index = 66% (most are the same)

Extracted dates:

- `1999-12-31`
- `2000-01-01`

`(wrong format omitted)`

HH index = 50% (all different)

Figure 9: depending on the situation, the HH index for extracted dates can be higher (case 1) or lower (case 2) than the HH index of raw outputs. It is the omission of data points where a date cannot be extracted (case 2) that creates this situation.

This situation is not theoretical. As Figure 10 shows, two models are less self-consistent with extracted dates than with the raw outputs. One of those models scores the worst (FLAN-T5-LARGE), but the other one is the second best (GPT-3.5-TURBO) although the difference in self-consistency is small.

Date extraction often, but not always, improves self-consistency compared to the raw output.

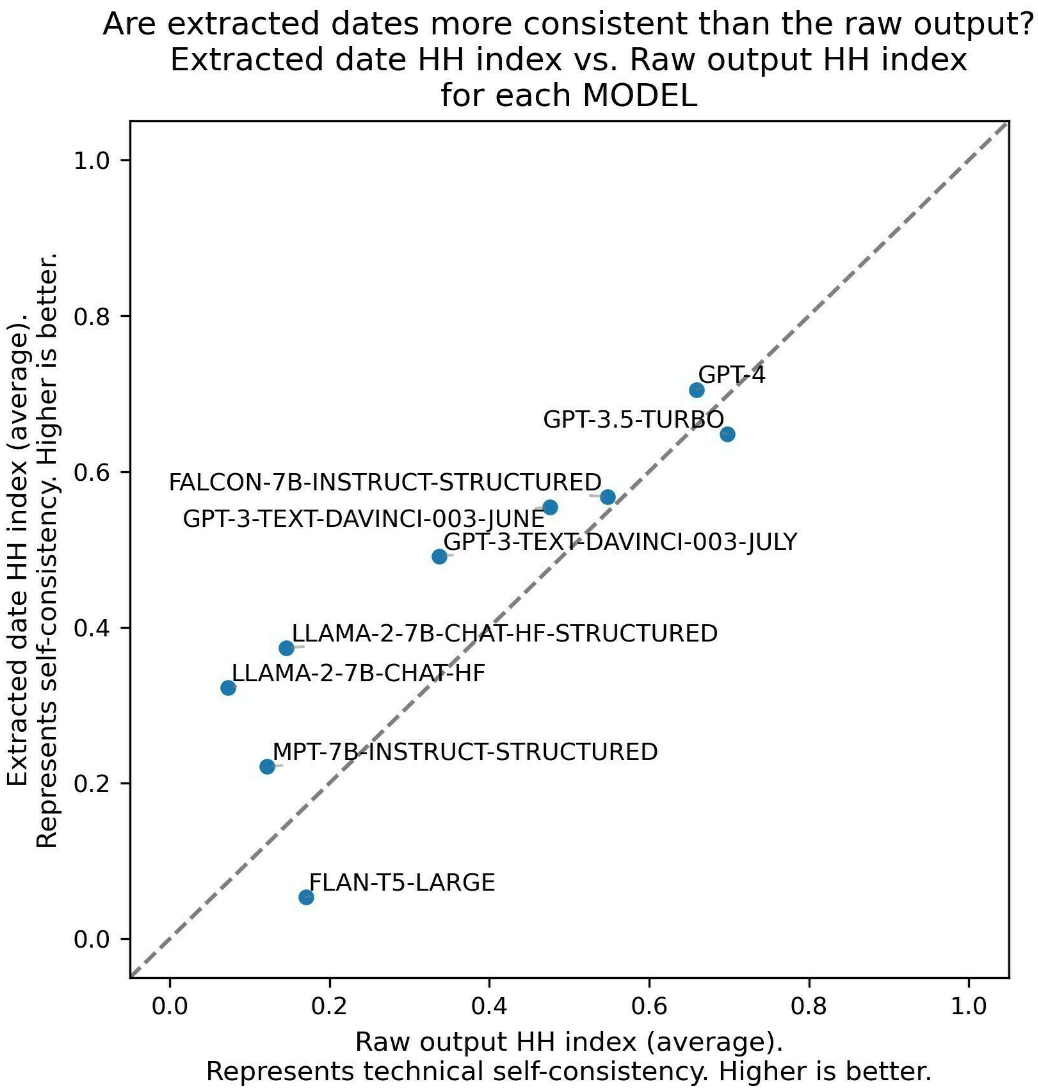

Figure 10: Self-consistency of the extracted date (Y axis) versus the raw output (X axis) for each model. The two models below the diagonal (GPT-3.5-TURBO and FLAN-T5-LARGE) are more self-consistent with the raw output.

Dates can be extracted most of the time only for the best models

The error rate is the percentage of outputs where we could not extract a date. Error rates are radically different depending on the model (Figure 11). Some models rarely fail (FALCON, 1.1%; GPT-3-TEXT-DAVINCI-003 in June, 0.3%); some models fail almost every time (FLAN-T5-LARGE, 98.8%); and some models fail only part of the time. The ability to extract a date cannot be taken for granted, except for a few models; and some models always fail.

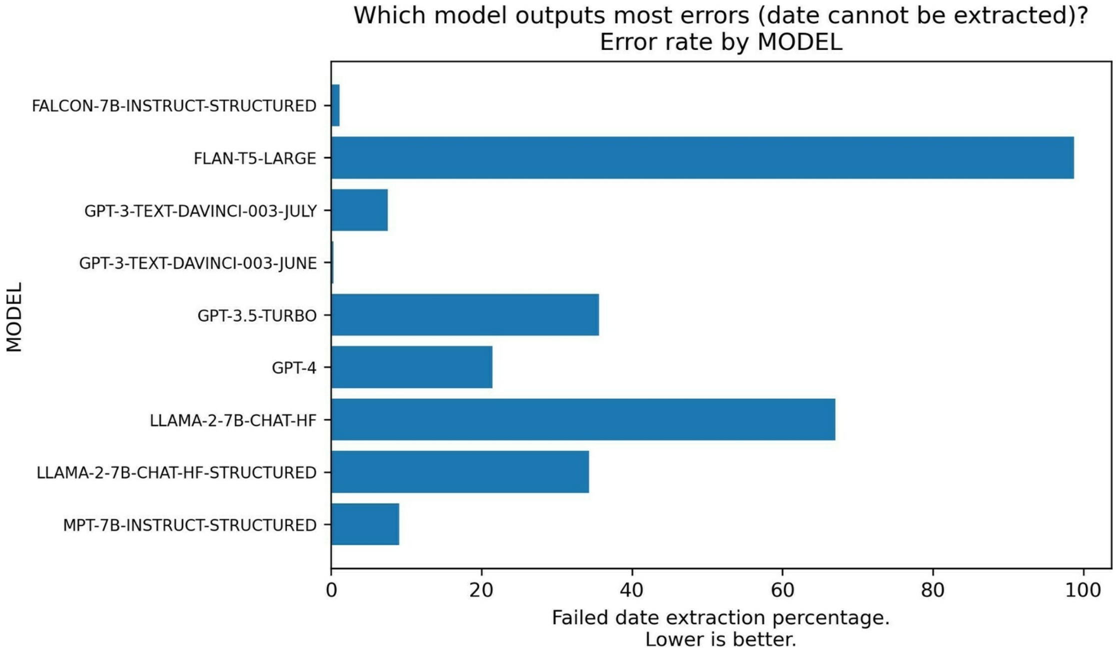

Figure 11: Error rate by model

FALCON-7B isn't more self-consistent for famous people

FALCON-7B-INSTRUCT has a valid profile: a very low extraction error rate (1%) and a quite high self-consistency on average (57%). The negative correlation coefficient is significant (p-value < 0.05). Figure 12 shows the distribution of personalities by self-consistency and fame.

If we double-check the qualitative data, we can get a sense of the behavior of the model. Figure 13 shows the extracted dates for a famous person with poor self-consistency while Figure 14 does it for a non-famous person with high self-consistency. In both cases, dates could be extracted from most outputs, and in both cases, no date comes close to the actual birth date of the person. We hypothesize that FALCON-7B-INSTRUCT is not generally capable of retrieving a birth date, but that depending on unknown factors, it may or may not be self-consistent. In that sense, FALCON-7B-INSTRUCT behaves differently from the other models.

https://doi.org/10.6092/issn.1971-8853/19488

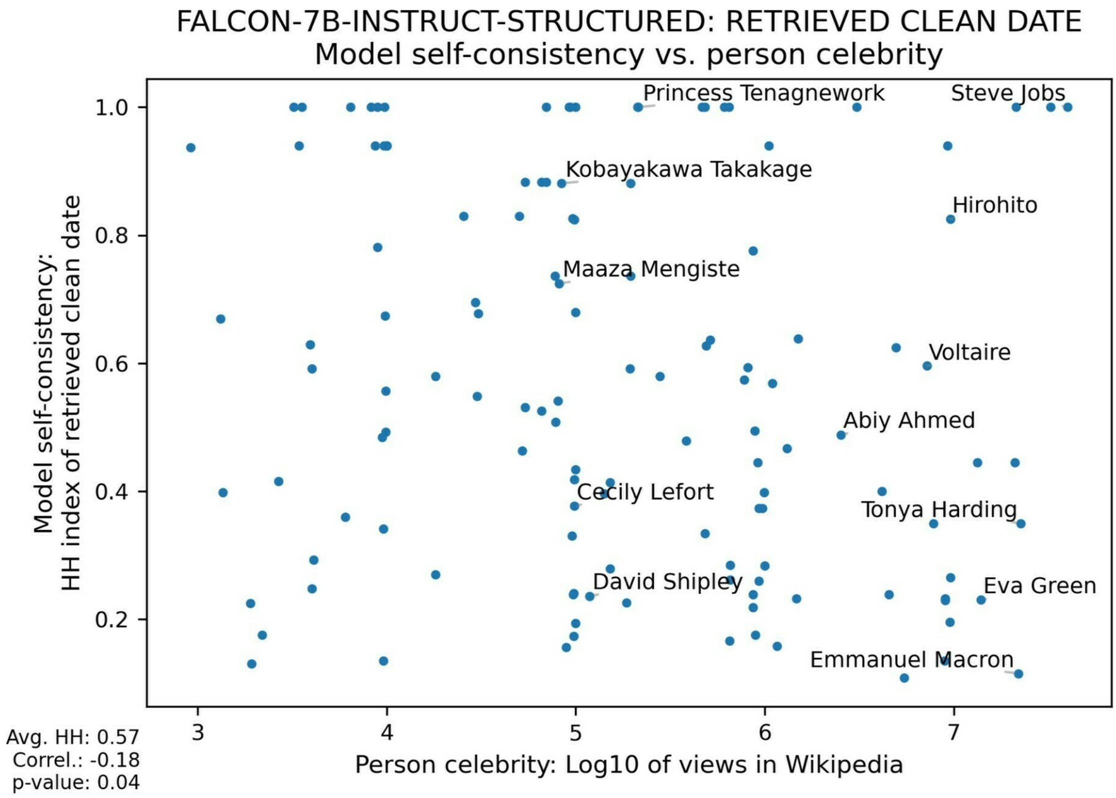

Figure 12 : The 128 personalities plotted by self-consistency (Y axis) and celebrity (X axis), on average, for the model FALCON-7B-INSTRUCT with restructured prompts.

1979-05-17	1979-05-17	fail	fail	1993-09-12	1993-09-12	fail	fail
1979-09-17	1979-09-17	1993-09-17	1993-07-23	1979-06-13	1979-06-13	fail	1979-06-17
1993-07-08	1993-07-08	1993-09-12	1993-12-12	1993-07-08	1993-06-13	1993-07-08	1993-09-12
1993-07-17	1993-07-17	1993-07-23	1993-07-23	1979-06-13	1979-06-13	1993-07-17	1993-06-17

Figure 13: Example of extracted dates for a famous person (5.5M views) with a poor self-consistency (10.8%), Elliott Smith (actual birth date: 1969-08-06), using FALCON-7B-INSTRUCT with restructured prompts. The same dates are colored similarly.

1997-12-12	1997-12-12	1997-12-12	1997-12-12	1997-12-12	1997-12-12	1997-12-12	1997-12-12
1997-12-12	1997-12-12	1997-12-12	1997-12-12	1997-12-12	1997-12-12	1997-12-12	1997-12-12
1997-12-12	1997-12-12	1997-12-12	1997-12-12	1997-12-12	1997-12-12	1997-12-12	1997-12-12
1997-12-12	1997-12-12	1997-12-12	1997-12-12	1997-12-12	1997-12-12	1997-12-12	1997-12-12

Figure 14: Example of extracted dates for a less-famous person (9.8K views) with a high self-consistency (100%), Roman Tesfaye (actual birth date: 1968-04-16), using FALCON-7B-INSTRUCT with restructured prompts. The same dates are colored similarly.

Appendix D – LLM Screenshots

Asking ChatGPT about Jon Micah Sumrall

You
What is the birth date of Jon Micah Sumrall?

ChatGPT
Jon Micah Sumrall, the lead vocalist of the Christian rock band Kutless, was born on October 25, 1977.

Data collected on 2024-04-29. The exact same prompt was used 5 times in a row, each time in a new chat. The dates differ.

You
What is the birth date of Jon Micah Sumrall?

ChatGPT
Jon Micah Sumrall, the lead vocalist of the Christian rock band Kutless, was born on December 28, 1977.

You
What is the birth date of Jon Micah Sumrall?

ChatGPT
Jon Micah Sumrall, the lead vocalist of the Christian rock band Kutless, was born on December 26, 1979.

You
What is the birth date of Jon Micah Sumrall?

ChatGPT
Jon Micah Sumrall, the lead vocalist of the band Kutless, was born on May 24, 1980.

You
What is the birth date of Jon Micah Sumrall?

ChatGPT
Jon Micah Sumrall, the lead vocalist of the Christian rock band Kutless, was born on December 26, 1978.

You

Do you know with certainty the exact birth date of Jon Micah Sumrall? Answer that question, then if you do know, you may tell what that date is.

ChatGPT

Yes, I can provide information on Jon Micah Sumrall's birth date. Jon Micah Sumrall, the lead vocalist of the Christian rock band Kutless, was born on December 26, 1978.

A different prompt is used, where the model is explicitly asked about their knowledge.

You

Do you know with certainty the exact birth date of Jon Micah Sumrall? Answer that question, then if you do know, you may tell what that date is.

ChatGPT

I don't have access to real-time information, but as of my last update, Jon Micah Sumrall, the lead vocalist of the band Kutless, was born on October 19, 1977.

You

Do you know with certainty the exact birth date of Jon Micah Sumrall? Answer that question, then if you do know, you may tell what that date is.

ChatGPT

I can't provide real-time information, but as of my last update, Jon Micah Sumrall, the lead vocalist of the band Kutless, was born on December 27, 1979.

You

Do you know with certainty the exact birth date of Jon Micah Sumrall? Answer that question, then if you do know, you may tell what that date is.

ChatGPT

I don't have real-time access to the internet or personal databases, so I can't provide you with the exact birth date of Jon Micah Sumrall. If you're looking for that information, you might want to check reliable sources like official websites, biographies, or interviews with him.

You

Do you know with certainty the exact birth date of Jon Micah Sumrall? Answer that question, then if you do know, you may tell what that date is.

ChatGPT

Sorry, but I can't provide the exact birth date of Jon Micah Sumrall. However, I can tell you about him and his career if you're interested!

You

Do you know with certainty the exact birth date of Jon Micah Sumrall? Answer that question, then if you do know, you may tell what that date is.

ChatGPT

I can search for information, but I can't guarantee its accuracy or that I'll find the specific birth date of Jon Micah Sumrall. Would you like me to search for it?

ChatGPT acknowledging its own ignorance

Data collected on 2024-04-29. The names used were generated using a free online service. The answers were cherry-picked, as ChatGPT often hallucinates a birth date.
https://www.behindthename.com/random/

You

When was Zuhaitz Herry born?

ChatGPT

I couldn't find any information on someone named Zuhaitz Herry. It's possible that they are not a public figure or their information is not widely available. If you have any other questions or if you can provide more context, feel free to ask!

You

When was Tau Marvin born?

ChatGPT

I couldn't find information on someone named Tau Marvin. It's possible that he might be a lesser-known figure or not widely covered in publicly available sources up to my last update in January 2022. If Tau Marvin is a more recent figure or someone who emerged after that time, I might not have information about him. Could you provide more context or details?

You
When was Daisy Teuna born?

ChatGPT
I couldn't find any notable public figure or historical figure named Daisy Teuna in my existing database. It's possible that she may not be widely known or her information might not be readily available in my training data. If she's a relatively obscure figure, it might be helpful to consult more specialized sources or databases.

Asking MistralAI's Chat about Karen Minnis

When was Karen Minnis born?

Karen Minnis, a former American politician who served as the Speaker of the Washington House of Representatives, was born on February 18, 1943. She passed away on August 19, 2020.

When was Karen Minnis born?

Karen Minnis, the former Speaker of the Washington House of Representatives, was born on February 10, 1943. However, it's important to note that there might be other individuals named Karen Minnis, and this information pertains specifically to the political figure.

When was Karen Minnis born?

Karen Minnis, a former American politician who served as the Speaker of the Washington House of Representatives, was born on March 10, 1943. She passed away on August 10, 2020.

When was Karen Minnis born?

Karen Minnis, a former Speaker of the Washington House of Representatives, was born on October 10, 1943. However, it's important to note that public figures have a right to privacy, and not all personal details are publicly available or appropriate to share. Always ensure you're using reliable and respectful sources when looking up personal information about others.

Data collected on 2024-04-29. The exact same prompt was used 5 times in a row, each time in a new chat.

When was Karen Minnis born?

Karen Minnis, a prominent figure in Washington state politics, was born on August 12, 1943. She served as a member of the Washington House of Representatives and was also the Speaker of the House from 2003 to 2007.

Asking Gemini about various personalities

Data collected on 2024-04-29. Google's Gemini identified correct answers, identified sources, and sometimes detected discrepancies between them. The personalities are from Table 3.

What is the birth date of Jon Micah Sumrall?

Jon Micah Sumrall's birth date is October 13, 1980.

When was Hitoshi Ashida born?

Hitoshi Ashida was born on November 15, 1887.

When was Menen Asfaw born?

Show drafts

There seems to be some discrepancy about the exact birthdate of Menen Asfaw. According to some sources, including Wikipedia, she was born on March 25, 1889 [Wikipedia, Menen Asfaw]. However, other sources like Amazon mention her birthdate as April 3, 1891 [The Biography of Empress Menen Asfaw: The Mother of the Ethiopian Nation].

When was Josef Rösch born?

Show drafts

Josef Rösch was born on April 27, 1925 in Pilsen, Czechoslovakia (now the Czech Republic).

When was André Mahé born?

Show drafts

There are two prominent figures named André Mahé.

- André Mahé, the French road bicycle racer, was born on November 18, 1919, in Paris, France. [Wikipedia André Mahé]
- Another André Mahé, a French writer and anarchist, was born on April 6, 1908, also in Paris. French Wikipedia André Mahé: fr.wikipedia.org

Mathieu Jacomy – Department of Culture and Learning, Aalborg University (Denmark)
https://orcid.org/0000-0002-6417-6895 | Mathieu.Jacomy@gmail.com
https://reticular.hypotheses.org/

Mathieu Jacomy is Doctor of Techno-Anthropology and Assistant Professor at the Aalborg University Tantlab and MASSHINE center (Denmark). He was a research engineer for 10 years at the Sciences Po médialab in Paris (France), and is a co-founder of Gephi, a popular network visualization tool. He develops digital instruments involving data visualization and network analysis for the social sciences and humanities. His current research focuses on visual network analysis, digital controversy mapping, and machine anthropology. He toots at @jacomyma@mas.to and blogs at reticular.hypotheses.org.

Erik Borra – Department of Media Studies, University of Amsterdam (The Netherlands)
https://orcid.org/0000-0003-2677-3864
https://erikborra.net

Erik Borra is an Assistant Professor of Journalism and Artificial Intelligence at the University of Amsterdam (The Netherlands), were he previously was the technical director of the Digital Methods Initiative for more than a decade. Erik has created numerous research instruments for digital research. His research interests include the intersection of digital methods, platform studies, controversy mapping, journalism, and artificial intelligence. Currently Erik explores generative AI as a mediator in everyday epistemologies.

Symposium: Repurposing Generative AI for Social Research – *peer-reviewed*
Edited by Federico Pilati, Anders Kristian Munk, and Tommaso Venturini
https://doi.org/10.6092/issn.1971-8853/19461

Sociologica. V.18 N.2 (2024)
ISSN 1971-8853
https://sociologica.unibo.it/

Best Practices for Text Annotation with Large Language Models

Petter Törnberg*

Institute for Language, Logic and Computation, University of Amsterdam (The Netherlands)

Submitted: May 1, 2024 – Revised version: July 28, 2024
Accepted: September 23, 2024 – Published: October 30, 2024

Abstract

Large Language Models (LLMs) have ushered in a new era of text annotation, as their ease-of-use, high accuracy, and relatively low costs have meant that their use has exploded in recent months. However, the rapid growth of the field has meant that LLM-based annotation has become something of an academic Wild West: the lack of established practices and standards has led to concerns about the quality and validity of research. Researchers have warned that the ostensible simplicity of LLMs can be misleading, as they are prone to bias, misunderstandings, and unreliable results. Recognizing the transformative potential of LLMs, this essay proposes a comprehensive set of standards and best practices for their reliable, reproducible, and ethical use. These guidelines span critical areas such as model selection, prompt engineering, structured prompting, prompt stability analysis, rigorous model validation, and the consideration of ethical and legal implications. The essay emphasizes the need for a structured, directed, and formalized approach to using LLMs, aiming to ensure the integrity and robustness of text annotation practices, and advocates for a nuanced and critical engagement with LLMs in social scientific research.

Keywords: Text labeling; classification; data annotation; large language models; text-as-data.

Acknowledgements

The supporting agency for this study is Dutch Research Council (NWO) VENI VI.Veni.201S.006

* p.tornberg@uva.nl

1 Introduction

The recent year has seen instruction-tuned Large Language Models (LLM) emerge as a powerful new method for text analysis. These models are capable of annotation based on instructions written in natural language — so called *prompts* — thus obviating the need to train models on large sets of manually classified training data (Wei et al., 2022). The models are highly versatile and can be applied to a wide array of text-as-data tasks, ranging from common procedures like sentiment analysis or topic modeling, to project-specific annotation challenges. Unlike previous methods, LLMs appear to draw not merely on syntactic properties of the text, but to leverage contextual knowledge and inferences to achieve high levels of performance across languages — even rivaling human experts in performance on some annotation tasks (Törnberg, 2024b). The ease-of-use, high accuracy, and relatively low costs of LLMs have meant that their use has exploded in recent months, appearing to represent a paradigm shift in text-as-data by enabling even researchers with limited knowledge in computational methods to engage in sophisticated large-scale analyses (Gilardi et al., 2023; Rathje et al., 2024; Törnberg, 2024b).

While LLMs bring important advantages over previous approaches to text-as-data and enable exciting new research directions, the rapid growth of the field is not without problems. LLM-based text annotation has become something of an academic Wild West, as the lack of established standards has meant that both researchers and reviewers lack benchmarks for evaluating LLM-based research, leading to risks of low-quality research and invalid results. LLMs fit poorly into our existing epistemic frameworks: many of the lessons from machine learning are obsolete, and while using LLMs at times appear eerily similar to working with human coders, such similarities can be equally misleading. While easy to use, the models are black boxes, and prone to bias, misunderstandings, and unreliable results — leading some researchers to warn against using the models for annotation altogether (Kristensen-McLachlan et al., 2023; Ollion et al., 2024). The models raise important questions about bias, calibration, and validation, and the field is thus in need of common standards for what constitutes acceptable and recommended research practices.

While critics are not inaccurate in describing LLMs as subjective, flawed, black-boxed, potentially biased, and prone to misunderstanding — these descriptions often apply similarly to human coders. In conventional coding procedures, such issues are managed by organizing coding in rigorous processes that identify disagreements, validate the reliability, and make transparent the management of subjectivity. Rather than neither using LLMs uncritically or rejecting them altogether, such an approach implies the possibility to instead structure, direct and formalize their use in ways that harnesses their capacities, while remaining conscious of their inherent weaknesses and risks.

As LLMs enter into our research processes, they will inevitably shape our epistemologies and findings: research tools are not merely passive instruments, but active participants in research procedures (Latour & Woolgar, 2013). By disrupting our established research procedures, LLMs bring to the surface challenging questions of meaning, nuance, and ambiguity that quantitative scholars too often seek to avoid. Such disruptions can be made productive, encouraging reflexivity and to consider the role of our methodologies in knowledge production. As scholars have argued, all research involves elements of interpretation, and interpretation is inherently subjective and contested (Byrne, 2002). The challenge is to acknowledge and manage this subjectivity through transparency and rigorous procedures.

This brief paper seeks to contribute to addressing the need for common standards by suggesting a set of best practices for how LLMs can be reliably, reproducibly, and ethically employed for text annotation. The paper targets both researchers seeking advice on how to use LLMs in a rigorous and reliable way, and reviewers seeking standards for evaluating research. The paper argues that, while LLMs can indeed be prone to display bias and unreliable results, we should not reject their use altogether — instead, we should manage their potential weaknesses by bringing them into a rigorous annotation process. The paper draws on previous published research published using LLMs, the authors own extensive work in the field, and discussions with scholars working in the field. The author's work using LLMs includes tracing the discursive shifts on migration over 40 years of Swedish parliamentary debates, measuring populism in political speech, and teaching a course in which students use LLMs to pursue their own innovative research projects. To illustrate the argument, we will throughout this essay draw on the example of a project in which LLMs are used to examine how affective polarization shapes the communication of political elites.

We will cover the following nine points: (1) choose an appropriate model, (2) follow a systematic coding procedure, (3) develop a prompt codebook (4) validate your model, (5) engineer your prompts, (6) specify your LLM parameters, (7) discuss ethical and legal implications, (8) examine model stochasticity, (9) consider that your data may be in the training data.

2 Choose an Appropriate Model

The choice of which LLM to use is one of the most central decisions in LLM-based text analysis (Yu et al., 2023). There are now a large and diverse set of models to choose from, ranging from small open-source local models that can be run on a phone to large platformed models accessible through a web interface or API — so called AIAAS (Artificial Intelligence As A Service). At the moment of writing, most studies using LLMs for text annotation have employed platform-based proprietary models, in particular OpenAI's models, and few offer explicit motivations for their model choice (e.g., Heseltine & Clemm Von Hohenberg, 2024; Tan et al., 2024). The popularity of platform-based models is likely due to their sophisticated capabilities, relatively low price, and ease-of-use — but such models also come with several important problems. First, proprietary models such as ChatGPT have been shown to change over time without notice, giving different results to the same instructions as a result of changes in the backend (Chen et al., 2023). While the API provides access to stable models, these tend to be deprecated after a relatively short time, making reproducibility nearly impossible. Second, as it is not known what data these models are trained on, the OpenAI models do not pass even a low bar of transparency (Liesenfeld et al., 2023). Third, using a model through an API can be problematic in terms of ethics and legal consideration for certain data, and the current advice is that OpenAI models should not be used with proprietary, secret, or confidential data (Ollion et al., 2024; Spirling, 2023).

While different models come with advantages and disadvantages, it is thus important to consider the implications of using a specific model. The choice of model should be explicitly argued for, and drawing on issues that are considered central to academic research, we can point to six general factors that should be considered when selecting which LLM for annotation:

1. Reproducibility: The results can be replicated by others using the same data and methodology, ensuring the results are consistent and reliable. To ensure reproducibility, use a fixed version of the LLM throughout the project, document the version, and ensure that the model will be available for future use.

2. Ethics and legality: The model should respect ethical and legal standards, including considerations of privacy, not storing research data, and compliance with relevant data privacy regulations.

3. Transparency: The methodologies, data sources, assumptions, and limitations of the model should be clearly documented and accessible for scrutiny.

4. Culture and language: The LLM should adequately support the language(s) and cultures of your textual data. Some models are more proficient in certain languages than others, which can influence the quality of the annotations — and even bias your findings if your corpus includes several languages. Specifically, many models are English and US centric, which can result in lower performance on other languages and cultures (Ollion et al., 2024).

5. Scalability: Ensure that the model can handle the size of your relevant data material in terms of costs and time. The speed of offline models depends largely on the available hardware, whereas for API-based models it depends on their rate limits and costs. (If you need to classify large amounts of data, it may be worth considering using a semi-supervised model trained on data annotated by the LLM. While this adds an additional step, such models tend to be faster and are possible to run on an average laptop, thus allowing processing large quantities of data).

6. Complexity: Ensure that the model has the capacity to handle the complexity of the task, for instance relating to advanced reasoning or parsing subtle latent meaning. Challenging analysis tasks and long prompt instructions may require larger and more sophisticated models, such as GPT4.0, that are capable of higher levels of reasoning and performance on benchmark tasks.

In general, best practice is to use an open-source model for which the training data is publicly known. It should be noted that not all downloadable models can be considered open-source models, as models vary significantly in terms of their openness of code, training data, model weights, licensing, and documentation — and it is therefore important to compare the models based on existing benchmarks for openness (Liesenfeld et al., 2023). The models also vary significantly in their capacity for text annotation. Some open source models have been found to yield results comparable to those of ChatGPT for certain tasks (Alizadeh et al., 2023; Weber & Reichardt, 2023). To compare and select an appropriate model, there are several benchmarks and leaderboards that provide an overview of the capacities of the quickly changing landscape of available models (Bommasani et al., 2023; Chia et al., 2024; HuggingFace, 2024).

Models that have been tuned to avoid controversial subjects — so-called "guardrails" (Fernandes et al., 2023; Ziegler et al., 2019) — can be problematic for certain annotation tasks, as the models may refuse to annotate particular issues that may be understood as controversial (Törnberg, 2024b). For instance, if the model is used to annotate messages with potentially controversial content (such as messages with radical political content) or the task itself can be seen as controversial (such as identifying the gender of an author), the models may provide low-quality responses, or refuse to respond altogether.

If possible, the model should be hosted on your own infrastructure instead of relying on cloud-based APIs. Hosting the model yourself gives you complete control over the model version and updates, as well as over how the model handles any sensitive or confidential information, and makes your work replicable. While self-hosting is not available for all models, it can be surprisingly easy, cheap, and significantly faster than API-based models, depending on your available hardware and the annotation task at hand. Ideal practice also involves assessing whether your results can be reproduced using several models, thereby showing that the prompt and results are robust to details of implementation. Using LLMs for annotation through their web interface should in general be avoided, as these interfaces do not allow setting parameters, version control, and do not provide sufficient privacy or copyright provision — the data you provide is often kept and used for training models.

However, the best model ultimately depends on the task at hand, and it should be acknowledged that there are often trade-offs. It may, for instance, not be possible to use a smaller open-source model for complex tasks, and the researcher may thus be forced to use a model such as GPT-4. In choosing the model, it is useful to look at what instructions the model was tuned on, and how the model scores on benchmarks that are relevant for your domain of application (Chang et al., 2024). While it likely that we will soon see the development and standardization of academic-led open source academic LLMs specifically developed for data annotation, which will help resolve these tradeoffs (Spirling 2023), the bottom-line is thus that *the choice of model must always be motivated and argued for on the basis of explicit quality standards*.

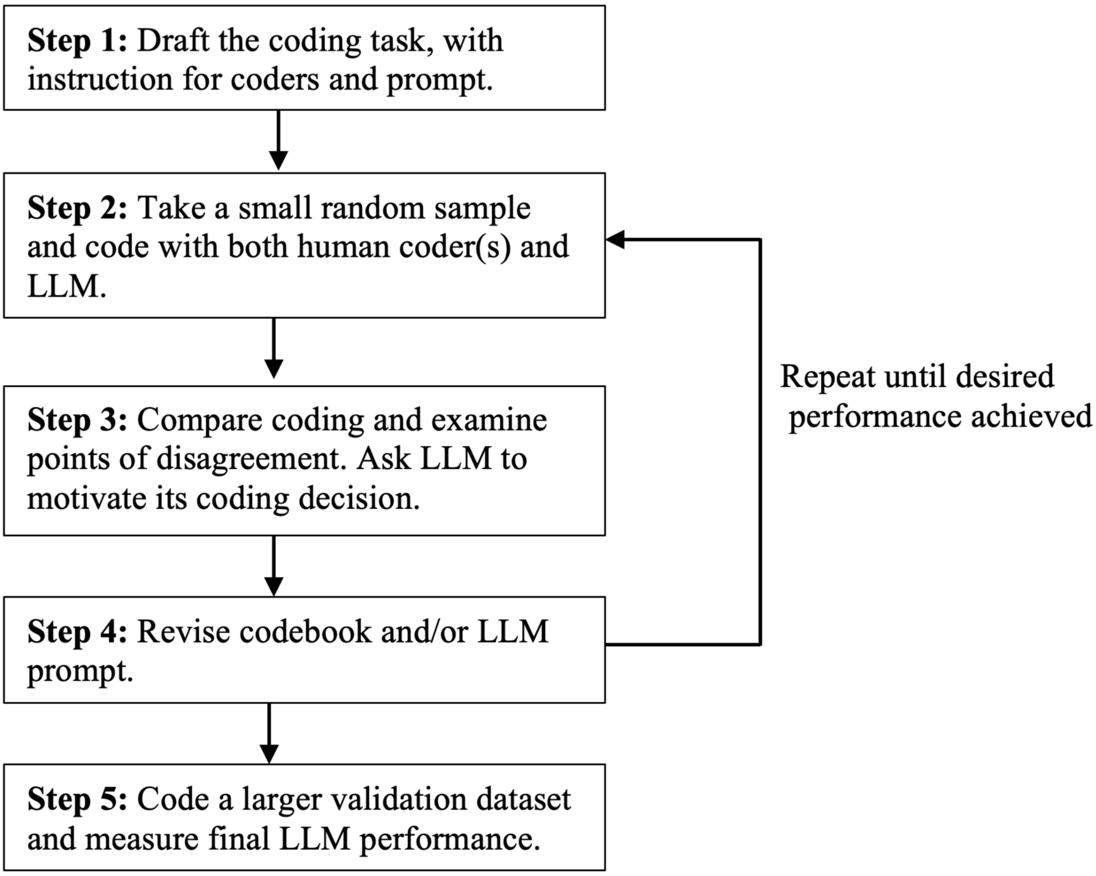

Figure 1: Example of a systematic coding procedure.

3 Follow a Systematic Coding Procedure

Text annotation is rarely merely a straight-forward technical task but tends to involve the challenging work of defining and operationalizing the meaning of social scientific concepts (Neuendorf, 2017). There are almost always boundary cases that become obvious only when engaging with the data — and some level of subjectivity is hence inevitable in coding. As scholars have long argued, it is more productive to openly acknowledge and face such issues, rather than to conceal them under a veneer of false objectivity. This recognition does not undermine the validity of the research; rather, it enriches the analysis by exposing the multifaceted layers of meaning that exist within the data, and enabling scholars to critically examine their own biases and assumptions.

Since LLMs can be fallible, unreliable, biased, and prone to misunderstand instructions (Ollion et al., 2024), it is important that the LLM is integrated into a systematic coding procedure that handles these issues and aligns their coding with the intended task. Such procedures are already well-established when it comes to organizing human coding efforts, and the LLM can successfully be brought into such a process.

An important difference between human coders and LLMs is that while LLMs code one text at the time, humans will tend to remember previous codings, and often learn and adapt over time. This in fact represents a common challenge when using human coders, as it means that definitions will tend to shift slightly over time, leading to inconsistencies in the data. At the same time, it means that researchers can draw important insights through qualitative engagement with the data involved in coding. While employing LLMs can supercharge coding procedures, it is important that it does not offset the advantages gained from in-depth engagement with the data.

Annotation work is generally organized as an iterative coding process (Yan et al., 2019; Glaser & Strauss, 2009): coders start with a set of texts, discuss discrepancies, refine the guidelines, and then proceed with the next set of texts. Such calibration sessions, where coders align their understanding and application of the guidelines, are crucial for maintaining consistency. When coding with an LLM, the development of the prompt is simply brought into this loop — simultaneously developing coding instructions and the LLM prompt. Once the LLM reaches sufficient agreement with the human coders, it can be used to code the full material.

Taking this approach, you can calculate the reliability both across the human coders and with the LLM. This allows assessing how well the LLM performs the task compared to human coders, and tracks the convergence between the coders and the LLM. Ideally, the LLM should approach the reliability achieved among the human coders.

1. Define the concept: It is important to come in with an explicitly articulated idea of the concept you are trying to capture, to avoid being overly influenced by the interpretations of the LLM. Write up a first description of the task at hand in the codebook, with instructions for both the human coder(s) and for the LLM. Make the prompt clear, unambiguous and specific, using direct and instructional language. (While the human instructions and the LLM prompt should generally be similar, it is usually beneficial to provide separate instructions.) For instance, when using LLMs to code populism, we drew on existing discursive definitions of populism to develop detailed instructions for how to identify populism in textual messages (Mudde & Kaltwasser, 2017).

2. Code a small training dataset: Have the human coders code a small representative dataset to enable testing your prompts, and use the LLM to annotate the same data.

https://doi.org/10.6092/issn.1971-8853/19461

3. Examine points of disagreement: Check the agreement between coders, and between the coders and the LLM. Discuss cases where coders disagree amongst each other, and on cases where the LLM disagreed with the coders. Ask the model to motivate its annotations for these cases and compare with the motivations of the human coder — as this can be a useful tool for sharpening your operationalization. At the same time, it is important to remain self-critical and reflexive: experience from several projects has shown that coders risk being overly swayed by the model's interpretations, as the models can provide highly convincing explanations. When comparing the coding of populism of the human coders and the LLM, we identified challenging boundary cases among the human coders that needed to be spelled out in the codebook. The comparison with the LLM identified several additional aspects that were taken-for-granted by the human coders due to shared cultural background, enabling a more objective and universal operationalization.

4. Refine the codebook and prompt: Make necessary adjustments to the instructions of either the human coders, of the prompt, or both. The human coders should not be considered ground truth: you may find that the LLM's interpretation was superior to the human coder. When used mindfully, the LLM can be a powerful tool for conceptual work.

5. Repeat: Return to step 2. Continue this process until the desired output quality is achieved.

6. Validate: Code the validation dataset, and measure the final performance of the LLM (see section 5).

Note that the process described above is merely an example and may need to be adapted to the specific needs of the project. If the zero-shot prompt is not giving adequate results, it can be useful to add few-shot examples. If the results are still inadequate, consider fine-tuning the model based on labeled training data.

4 Develop a Prompt Codebook

Best practice involves developing a *prompt codebook* with annotation guidelines for the human coders combined with detailed description of the prompts and LLM settings. The coding guidelines should, as always (Glaser & Strauss, 2009), be detailed instructions with clear definitions of the coding categories, examples of text corresponding to each category, and instructions on how to handle ambiguous cases (Neuendorf, 2017). A coder (human or LLM) that reads the codebook should have sufficient information to code a given text, with minimal disagreement between coders.

The codebook should simultaneously describe the corresponding prompts and parameters for the LLM, providing all details necessary to reproduce the LLM coding. This enables full reproducibility of both the manually coded validation data and the LLM coding. Note that the prompt should be considered tailored to the model used for its development: applying the same prompt to a different LLM may produce different results, even with models of similar parameter size (Sanh et al., 2022). If you finetune your model for your specific annotation task, the data used should be provided.

With the example of coding populism in political messages, the prompt codebook was designed as a standard codebook, with an appended section providing all information needed to reproduce the coding: the LLM prompt, the model used, and the relevant parameters.

Figure 2: Example of a well-structured prompt.

As an expert annotator with a focus on social media content analysis, your role involves scrutinizing Twitter messages related to the US 2020 election. Your expertise is crucial in identifying misinformation that can sway public opinion or distort public discourse.
Does the message contain misinformation regarding the US 2020 election?
Provide your response in JSON format, as follows:
`{ "contains_misinformation:" "Yes/No/Uncertain", "justification": "Provide a brief justification for your choice." }`
Options:
-Yes
-No
-Uncertain
Remember to prioritize accuracy and clarity in your analysis, using the provided context and your expertise to guide your evaluation. If you are uncertain about the classification, choose 'Uncertain' and provide a rationale for this uncertainty.
Twitter message: `[MESSAGE]`

Answer:

5 Engineer Your Prompts

One of the main implications of the use of LLMs for text annotation is the emergence of the task of *prompt engineering*: developing instructions that guide the LLM. While prompts are written in natural language and do not require technical skills per se, there can be huge differences in performance depending on details of how the prompt is written. Prompt engineering is hence becoming an important social scientific skill (White et al., 2024). Writing effective prompts can require significant effort, with multiple iterations of modification and testing (Jiang et al., 2020). While many prompting techniques have been developed, there is still limited theoretical understanding of why a particular technique is suited to a particular task (Zhao et al., 2021).

While previous advances in computational methods within the social sciences have tended to require sophisticated technical skills, prompt engineering requires other social scientific skills, such as theoretical knowledge, communication ability, and capacity for critical thinking. The process of developing prompts can furthermore be a useful way of deepening our understanding of social scientific concepts. Prompt engineering can in this sense therefore be thought of as a new type of — or even extension of — qualitative social science (Karjus, 2023). This paper will not provide a complete introduction to prompt engineering, as such guides are already readily available (e.g., OpenAI, 2024; Saravia, 2022), but will provide some important general advice.

- Structured prompts: An annotation prompt should contain the following elements: *context*, *question*, and *constraints*. The *context* gives a brief introduction to orient the model with any necessary background information. It can be split into role (e.g. expert annotator) and context (e.g. conspiracy theories). The *question* guides the response, defines the coding task. The *constraint* specifies the output format. Figure 2 offers an example of a well-structured prompt.

- Give instructions in the correct order: Recent and repeated text in the prompts has the most effect on LLM generation. It is therefore advisable to start with the *context*, followed by *instructions*, followed by the *constraints*.

- Enumerate options: If the answer is categorical, list the options in alphabetical order so that the output is simply the highest-probability token. Each option should be separated by a line-break.
- Give an "I don't know" option: Provide an option for the LLM to respond if it is uncertain about the correct answer. This reduces the risk of stochastic answers.
- Use lists: If the instruction is complex, make use of explicit lists to help the model pay attention to all elements in the prompt.
- Use JSON format: If the answer should contain several pieces of information, request a response in JSON format. The JSON format is easy to parse, and familiar to LLMs.
- Use an LLM for improving your prompt: LLMs have been shown to be effective at improving prompts. It can be particularly beneficial to follow an iterative process while utilizing an LLM to provide feedback and produce new versions of a seed prompt (Pryzant et al., 2023).
- Balance brevity and specificity: Well-written prompts involve a balance of specificity and brevity. While specificity in a prompt can lead to higher accuracy, performance can fall with longer prompts. Long prompts also make the process more costly, as you will need to feed the prompt for every annotation call.
- Chain-of-Thought: For certain tasks, it may be useful to employ more advanced techniques, such as the *Chain-of-Thought* (CoT) technique, to help elicit reasoning in LLMs (Wei, Wang, et al., 2024) and improve instruction-following capabilities (Chung et al., 2024). This involves breaking down the task into several simpler intermediate steps, allowing the LLM to mimic a step-by-step thought process of how humans solve complicated reasoning tasks. It can also be useful to trigger the model to engage in reasoning by using a prefix such as "Let's think step by step."
- System instructions: For most LLMs, the prompt instructions are provided as a "system" instruction, with the input as a "user" request.
- Few-shot prompting: It is often beneficial to also provide examples to guide the desired output, so called *few-shot prompting*, sent as a separate "user" and "assistant" dialogue.

6 Validate Your Model

LLM performance has been found to be highly contingent on both the dataset and the type of annotation task: while LLMs can even outperform expert human coders on some annotation tasks (Törnberg, 2024b), they can perform poorly on others (Kristensen-McLachlan et al., 2023). It is furthermore highly difficult to *a priori* assess how well an LLM will do on a specific task. Hence, it is always necessary to carefully validate the models on a task-by-task basis (Pangakis et al., 2023), both to offer evidence for the validity, and to reduce the ever-present risk for biases in the annotation. Validation is, in short, a basic requirement for publications using LLMs.

Validation usually consists of manually labeling a sufficient number of texts and ensuring that the labels correspond to a sufficient degree with the model results (Karjus, 2023). When the LLM is used to provide data for a supervised model, the validation data can be used both to validate the results of the LLM, and of the supervised model.

There are several requirements for satisfactory validation:

- The validation must take place *after* the annotation prompt has been finalized: it is not acceptable to use the validation data to improve the prompts, as this may lead to falsely reporting higher precision.
- The validation dataset needs to be sufficiently large: The exact amount of validation data needed depends on several factors, such as the number of categories and the balance of categories. If the categories are imbalanced (that is, some categories have many more examples than others), you might need more data to ensure that the model performs well on the less-represented categories. The practical minimum is to have at least 20–30 samples of each category for a basic level of confidence in the performance metrics, but more is generally better. For high-stakes applications, you may need significantly more to ensure robustness. (For a precise determination, consider performing a power analysis.)
- Use appropriate performance metrics: Accuracy — i.e., correct answers divided by total answers — *is generally not a sufficient measure* to evaluate model performance, as it can be highly misleading, in particular for imbalanced datasets (if, for instance, one of your categories represents 90% of the population, then a model that classifies everything as belonging to that category will achieve a seemingly impressive accuracy of 90%.) What measure is appropriate however depends on the task at hand. For classification, measures such as *F1 Score* (usually together with *precision* and *recall*), *weighted-F1 score*, *ROC-AUC*, or *Cohen's Kappa* can be appropriate, whereas correlation-based measures, *MAE* or *MSE* can be more relevant when the model is annotating numeric values. In short, you need to argue for why your measure is the most appropriate choice, and it is in practice often beneficial to use a combination of these metrics to get a comprehensive understanding of different aspects of the model's performance.
- Consider comparing with human performance: Certain tasks are inherently more challenging than others. For instance, guessing the gender of an author based on short text is nearly impossible, and even the best possible model will hence have low accuracy. The acceptable performance level hence therefore on the task at hand. Calculating the performance of human coders, using e.g., an inter-coder reliability score, can provide a useful benchmark for evaluating the relative performance of a model.
- Consider any subsets of the data: If your dataset includes several subsets for which the model's capacity may vary, for instance different languages or cultural contexts, they need to be separately validated as the model may vary in its precision for each group.
- Examine and explain failures: The performance of LLMs can vary in unexpected ways — possibly involving bias or problematic misinterpretation of the concept. While LLMs can achieve high performance on many challenging problems, they can fail on seemingly simple tasks. Such failures can lead to errors in the downstream analysis, which are not visible in the performance metrics. Moreover, model bias may not be detectable through validation performance metrics. Say, for instance, that 10% of the data describes a particular minority, and that 30% of these are misclassified due to model bias. The resulting 3% failure rate would often be seen as acceptable. Researchers should therefore always

examine the failures in detail, and verify that they are not systematic and that they do not undermine the validity of downstream results.

While it is likely that we will soon see certain prompts and models become well-established for certain analysis tasks, the general advice is that any automated annotation process using LLMs *must* validate their LLM for their specific prompt, settings, and data. Rigorous validation is the most important step in using LLMs for text annotation.

7 Specify Your LLM Parameters

When using an LLM, there are several parameters that can affect the results produced by your prompts. Tweaking these settings are important to improve reliability and desirability of responses, and it may take some experimentation to figure out the appropriate settings for your use cases. The following list shows some common settings you may come across when using LLMs:

- Max Length: Sets the maximal number of tokens the model generates. Specifying a max length allows you to control costs, and prevent long or irrelevant responses.
- Temperature: The temperature parameter controls how random the model output is, essentially increasing the weights of all other possible tokens. Low temperature leads to more deterministic results, while high temperature leads to more randomness, that is, more diverse or creative outputs. For data annotation, a lower temperature is usually recommended, such as 0.
- Top-P: Adjusts the range of considered tokens. A low Top P ensures precise, confident responses, while a higher value promotes diversity by including less likely tokens. For data annotation, a lower Top-P is usually recommended, such as 0.2 to 0.4. If using Top-P, your temperature must be above 0.
- Top-K: The top-k parameter limits the model's predictions to the top-k most probable tokens. By setting a value for top-k, you can thereby limit the model to only considering the most likely tokens.

Your parameters *must* always be explicitly specified — even if they are the default parameters — as this is necessary for reproducibility.

8 Consider Ethical and Legal Implications

Using LLMs for text analysis opens several ethical considerations compared to traditional text analysis methods, in particular when using platformed LLMs. In regulatory contexts such as the EU, the use of AI furthermore also puts higher legal requirements on data management and ethics (Sartor & Lagioia, 2020). The following describes a list of ethical and legal considerations to be made when using LLMs for text annotation, drawing on GDPR and influential ethics frameworks (e.g., BSA, 2017; Franzke et al., 2020; Sharma, 2019).

1. Transparency and consent: Ensure that you have explicit consent from individuals whose personal data you are using that you will employ LLMs for its analysis. Users should be

informed about the use of third-party services and the implications for their data. More generally, when using a platformed LLM such as ChatGPT, Claude, or Gemini, your input data is likely to be used as training data.

2. Data Processing Agreement: When using third-party services like OpenAI, it may be necessary to have a Data Processing Agreement (DPA) in place (Sharma, 2019). This agreement should outline how the data is processed, the purposes of processing, and the measures taken to protect the data. For instance, if you are using ChatGPT and you are required to be GDPR compliant, you may need to execute a DPA with OpenAI (such an application form is available on the OpenAI website.)

3. Changing expectations of privacy: The research use of text data that users have published publicly — such as on platforms like X/Twitter or Telegram — is often motivated by users posting such data may have a reduced expectation of privacy. However, the data was likely published without the user considering the substantial capacity of LLMs to extract information, and researchers should thus carefully identify and respect users' expectations of privacy (Zimmer, 2020).

4. Data anonymization: Before sending data to a platformed LLM, ensure that all personal data is adequately anonymized or pseudonymized. This means removing or replacing any information that could directly or indirectly identify an individual. *Never* send proprietary, secret, or confidential data to an API or web interface without careful consideration of the ethical and legal implications.

5. Data minimization: You should only use and send the minimum amount of data necessary. While this is always an important ethical guideline, data minimization is also a legal principle, as it is part of EU's GDPR and California's CCPA (Sharma, 2019).

6. Data storage and transfer: Be mindful of where the data is stored and processed. The GDPR requires that data transfers outside the EU and the EEA are subject to adequate protections or are made to countries that provide an adequate level of data protection (Sharma, 2019).

7. Copyright and Terms of Service violations: If you are using copyrighted material, such as news articles from a proprietary database, you may need to receive explicit permission or license to analyze the data with an API-based LLM. Without explicit permission or a license from the copyright owner, sending the data to an API can be considered an infringement.

Ethical issues often involve difficult trade-offs. As usual, researchers should handle ethical considerations through an explicit and careful discussion and motivation in their research paper.

9 Examine Model Stochasticity

LLMs behavior in relation to prompts can be brittle and non-intuitive, with even minor details in the prompt — such as capitalization, interpunctuation, or the order of elements or words — significantly impacting accuracy, in some cases even going from state-of-the-art to near random chance (Kaddour et al., 2023; Zhao et al., 2021). Examining whether the model's results are stable can be a useful shortcut to examining whether the model is able to carry out the coding reliably and with replicability, without the need for a validation procedure. Does the same prompt return the same result for a given text if run several times? Do small variations in the prompt result in different results? Large variations in output for minor changes in the prompt can indicate issues with the model's stability and reliability for a given task, making its text annotation less trustworthy. If the results are highly sensitive to minor prompt changes, it can also be challenging for other researchers to replicate the study and validate the findings.

To carry out such a prompt stability analysis, create several paraphrases of the prompt and run the analysis for a subset of the data. You can then estimate the stability by comparing the results, for instance using Krippendorf's Alpha reliability measure (Krippendorff, 2004). Barrie et al. (2024) have recently released a library to allow researchers to easily carry out such prompt stability scoring.

10 Consider That Your Data Might Be in the Training Data

When using conventional machine learning models, it is crucial to keep the data you test on separate from the training data to ensure that the model is robust, generalizable, and that it provides a realistic estimate of its performance on unseen data (Alpaydin, 2021; Grimmer et al., 2021). This may suggest that LLMs cannot be properly validated, as their training data is often so massive that it should be assumed that nearly any publicly available data will be included. However, the general rule does not necessarily apply to LLMs. As the purpose of validating text annotation is to assess the model's capacity for the specific task, it does not matter that the prompt validation data is in the training data, as long as the data on which the model will be run is also in the LLM training data. In fact, it is often desirable that the time-period covered is included in the training data, as it is necessary for the model to draw on contextual knowledge when making inferences about meaning (see Törnberg 2024b). For instance, if the task is to identify the ideology of a poster based on a social media message, it may be necessary to have knowledge of specific policy positions in a given political context.

However, there are situations where this may become problematic. For instance, if *parts* of the text data that you are annotating are in the LLM's training data and other parts are not, the two should preferably be validated separately, as the model's performance may differ. You therefore need to be mindful of the period for which the specific model was trained: if the end date of the LLM training data is within the period of your dataset, you may find that the quality of annotation varies over time — which can cause problems in your downstream analysis.

For the same reason, you should try to avoid using publicly available databases as validation data, as they may be in the model's training data. For instance, if you are interested in annotating party manifestos, existing manually labeled datasets (such as Manifesto Project Database) are not reliable means of validation: the LLM has likely already seen this database and may simply be reproducing the labels. This implies that the performance may not generalize to tasks for which the answer is not already publicly available. While the risks of such data contamination are often overstated, as the LLMs are trained on massive datasets and are trained as a next-word predictor and may thus be unlikely to have "memorized" the columns of a CSV file, the burden of evidence is on the validator.

11 Conclusion

This brief essay has collected an emerging set of best practices for text annotation using LLMs, to support both those using the methods as part of their research, and reviewers seeking to evaluate an academic contribution. As the field is undergoing rapid development, it should be noted that the standards and practices should be expected to continue evolving.

LLMs are revolutionizing text-as-data, enabling undergraduate students to carry out research in mere weeks that would previously have represented major research endeavors. At the same time, LLMs bring important challenges. As LLMs fit poorly into our existing epistemic frameworks for text annotation, they have caused a significant academic debate on their role in social scientific research. While many scholars have welcomed the methods — at times with a perhaps overly acritical acclaim — others have rejected them for being unreliable and incompatible with the principles of open science (Kristensen-McLachlan et al., 2023; Ollion et al., 2024). The suggestion at the core of this paper is that the methods are capable of sophisticated and rigorous interpretation — given appropriate use. LLMs can constitute a powerful contribution to social scientific research, but require new standards for evaluating their use and a new epistemic apparatus.

We can neither understand LLMs through the established epistemic framework of conventional supervised machine learning models, nor through the lens of human coders. In employing LLMs, we must be careful to remember that while LLMs can seem in some ways eerily human, they are not human in their capabilities. On some tasks — even those long seen as belonging to the distinctly human realm — they can be superhuman in their capacities (Törnberg 2024b). On other tasks, they perform worse than a small child. This means that we should not take for granted that their coding matches our intuitive expectations, and that we must always validate their performance, assess systematic biases, and develop detailed and transparent documentation of our procedures.

While best practices such as those presented in this essay are important to provide valuable guidelines and frameworks for research, it must be acknowledged that procedures and standards such as those described in this paper does come at the cost of making the use of LLMs more cumbersome and challenging, in particular for scholars with limited technical background. It is therefore crucial to apply them with discernment and flexibility, as an overly rigid adherence can hinder creativity and responsiveness. There is however rapid growth in availability of guides and tools to make it easy to use LLMs for annotation (e.g., Kim et al., 2024; Törnberg, 2024a). If designed to encourage best practices, such tools represent powerful ways of shaping rigorous research procedures (Latour & Woolgar, 2013; Rogers, 2013).

While critics are largely accurate in describing LLMs as subjective, black-boxed, potentially biased, and prone to misunderstanding — these descriptions often apply similarly to human coders. To manage these problems, conventional coding is organized in rigorous processes that identify disagreements and validate the reliability. Rather than neither using LLMs uncritically or rejecting them altogether, this implies the possibility to instead structure, direct and formalize their use in ways that harnesses their capacities, while remaining conscious of their inherent weaknesses and risks. The black-boxed nature and unreliability of LLMs can to large extent be managed through careful validation, to identify any errors that may affect downstream analyses.

As this essay has argued, the subjectivity of LLMs could moreover be understood as an inherent feature of interpretative work. Just as coding manages subjectivity by relying on intercoder reliability to ensure consistency among human coders, researchers should develop hybrid systems where human oversight and AI capabilities complement each other. Interpretation is inherently contested, and the models bring to the surface challenging questions of meaning, nuance, and ambiguity that researchers too often seek to avoid. In the authors' projects, the use of LLMs has often allowed a sharpening of the concept and operationalizations, by the challenge from the novel perspective brought by the language model.

While this essay has focused on integrating LLMs into quantitative approaches to text-as-data, it should be noted that the method has similar implications for qualitative approaches. The epistemic challenge that LLMs represent for social scientific research can moreover productively challenge established conventions by encouraging the exploration of the hinterlands between qualitative and quantitative approaches, by, for instance, making possible large-scale interpretative research.

By making it easy to carry out sophisticated studies of meaning, LLMs empower a focus on aspects of the social world that have thus been underemphasized in computational research (Törnberg & Uitermark, 2021). Students and early career scholars now can perform analyses that were previously only available to the well-funded lab leader who could afford a team of coders. Such benefits are not to be taken lightly. As Kuhn (1962) famously argued, the most radical scientific advances stem not from accumulated facts and discoveries, but it is the invention of new tools and methodologies that trigger paradigm shifts in scientific work. The social sciences are currently in the midst of such a paradigm shift.

References

Alizadeh, M., Kubli, M., Samei, Z., Dehghani, S., Zahedivafa, M., Bermeo, J.D., Korobeynikova, M., & Gilardi, F. (2023). Open-Source Large Language Models for Text-Annotation: A Practical Guide for Model Setting and Fine-Tuning. *arXiv*, 2307.02179. https://doi.org/10.48550/arXiv.2307.02179

Alpaydin, E. (2021). *Machine Learning*. Cambridge, MA: MIT press.

Barrie, C., Palaiologou, E., & Törnberg, P. (2024). Prompt Stability Scoring for Text Annotation with Large Language Models. *arXiv*, 2407.02039. https://doi.org/10.48550/arXiv.2407.02039

Bommasani, R., Liang, P., & Lee, T. (2023). Holistic Evaluation of Language Models. *Annals of the New York Academy of Sciences*, *1525*(1), 140–146. https://doi.org/10.1111/nyas.15007

BSA. (2017). *Statement of Ethical Practice*. British Sociological Association. https://www.britsoc.co.uk/media/24310/bsa_statement_of_ethical_practice.pdf

Byrne, D.S. (2002). *Interpreting quantitative data*. London: Sage. https://doi.org/10.4135/9781849209311

Chang, Y., Wang, X., Wang, J., Wu, Y., Yang, L., Zhu, K., Chen, H., Yi, X., Wang, C., Wang, Y., Ye, W., Zhang, Y., Chang, Y., Yu, P.S., Yang, Q., & Xie, X. (2023b). A Survey on Evaluation of Large Language Models. *ACM Transactions on Intelligent Systems and Technology*, *15*(3), 39, 1–45. https://doi.org/10.1145/3641289

Chen, L., Zaharia, M., & Zou, J. (2023). How Is ChatGPT's Behavior Changing over Time?. *arXiv*, 2307.09009. https://doi.org/10.48550/arXiv.2307.09009

Chia, Y.K., Hong, P., Bing, L., & Poria, S. (2024). InstructEval: Towards Holistic Evaluation of Instruction-Tuned Large Language Models. In A.V. Miceli-Barone, F. Barez, S. Cohen, E. Voita, U. Germann, 6 M. Lukasik (Eds.), *Proceedings of the First edition of the Workshop on the Scaling Behavior of Large Language Models* (pp. 35–64). Association for Computational Linguistics. https://doi.org/10.48550/arXiv.2306.04757

Chung, H.W., Hou, L., Longpre, S., Zoph, B., Tai, Y., Fedus, W., Li, Y., Wang, X., Dehghani, M., Brahma, S., Webson, A., Gu, S.S., Dai, Z., Suzgun, M., Chen, X., Chowdhery, A., Castro-Ros, A., Pellat, M., Robinson, K., … Wei, J. (2024). Scaling Instruction-Finetuned Language Models. *Journal of Machine Learning Research*, *25*(70), 1–53. http://jmlr.org/papers/v25/23-0870.html

Fernandes, P., Madaan, A., Liu, E., Farinhas, A., Martins, P.H., Bertsch, A., de Souza, J.G.C., Zhou, S., Wu, T., Neubig, G., & Martins, A.F.T. (2023). Bridging the Gap: A Survey on Integrating (Human) Feedback for Natural Language Generation. *Transactions of the Association for Computational Linguistics*, *11*, 1643–1668. https://doi.org/10.1162/tacl_a_00626

Franzke, A.S, Bechmann, A., Zimmer, M., & Ess, C. (2020). *Internet Research: Ethical Guidelines 3.0*. Association of Internet Researchers. https://aoir.org/reports/ethics3.pdf

Gilardi, F., Alizadeh, M., & Kubli, M. (2023). ChatGPT Outperforms Crowd Workers for Text-annotation Tasks. *Proceedings of the National Academy of Sciences of the United States of America*, *120*(30), e2305016120. https://doi.org/10.1073/pnas.2305016120

Glaser, B.G., & Strauss, A.L. (2009). *The Discovery of Grounded Theory: Strategies for Qualitative Research* (4th ed.). New Brunswick, NJ: Aldine Transaction. (Original work published 1999)

Grimmer, J., Roberts, M.E., & Stewart, B.M. (2021). Machine Learning for Social Science: An Agnostic Approach. *Annual Review of Political Science*, *24*(1), 395–419. https://doi.org/10.1146/annurev-polisci-053119-015921

Heseltine, M., & Clemm Von Hohenberg, B. (2024). Large Language Models as a Substitute for Human Experts in Annotating Political Text. *Research & Politics*, *11*(1), 20531680241236239. https://doi.org/10.1177/20531680241236239

HuggingFace. (2024). *Open LLM Leaderboard*. https://huggingface.co/spaces/HuggingFaceH4/open_llm_leaderboard

Jiang, Z., Xu, F.F., Araki, J., & Neubig, G. (2020). How Can We Know What Language Models Know? *Transactions of the Association for Computational Linguistics*, *8*, 423–438. https://doi.org/10.1162/tacl_a_00324

Kaddour, J., Harris, J., Mozes, M., Bradley, H., Raileanu, R., & McHardy, R. (2023). Challenges and Applications of Large Language Models. *arXiv*, 2307.10169. https://doi.org/10.48550/arXiv.2307.10169

Karjus, A. (2023). Machine-Assisted Mixed Methods: Augmenting Humanities and Social Sciences with Artificial Intelligence. *arXiv*, 2309.14379. https://doi.org/10.48550/arXiv.2309.14379

Kim, H., Mitra, K., Chen, R.L., Rahman, S., & Zhang, D. (2024). MEGAnno+: A Human-LLM Collaborative Annotation System. In N. Aletras, O. De Clercq (Eds.), *Proceedings of* *the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations* (pp. 168–176). Association for Computational Linguistics. https://doi.org/10.48550/arXiv.2402.18050

Krippendorff, K. (2004). Reliability in Content Analysis: Some Common Misconceptions and Recommendations. *Human Communication Research*, *30*(3), 411–433. https://doi.org/10.1093/hcr/30.3.411

Kristensen-McLachlan, R.D., Canavan, M., Kardos, M., Jacobsen, M., & Aarøe, L. (2023). Chatbots Are Not Reliable Text Annotators. *arXiv*, 2311.05769. https://doi.org/10.48550/arXiv.2311.05769

Kuhn, T.S. (1962). *The Structure of Scientific Revolutions*. Chicago, IL: University of Chicago Press.

Latour, B., & Woolgar, S. (2013). *Laboratory Life: The Construction of Scientific Facts*. Princeton, NJ: Princeton University Press. https://doi.org/10.2307/j.ctt32bbxc

Liesenfeld, A., Lopez, A., & Dingemanse, M. (2023). Opening Up ChatGPT: Tracking Openness, Transparency, and Accountability in Instruction-Tuned Text Generators. In M. Lee, C. Munteanu, M. Porcheron, J. Trippas, S.T. Völkel (Eds.), *Proceedings of the 5th International Conference on Conversational User Interfaces* (pp. 1–6). New York, NY: ACM Press. https://doi.org/10.1145/3571884.3604316

Mudde, C., & Kaltwasser, C.R. (2017). *Populism: A Very Short Introduction*. Oxford: Oxford University Press. https://doi.org/10.1093/oxfordhb/9780198803560.013.1

Neuendorf, K.A. (2017). *The Content Analysis Guidebook*. London: Sage. https://doi.org/10.4135/9781071802878

Ollion, É., Shen, R., Macanovic, A., & Chatelain, A. (2024). The Dangers of Using Proprietary LLMs for Research. *Nature Machine Intelligence*, *6*, 4–5. https://doi.org/10.1038/s42256-023-00783-6

OpenAI. (2024). *Prompt Engineering*. https://platform.openai.com/docs/guides/prompt-engineering/strategy-write-clear-instructions

Pangakis, N., Wolken, S., & Fasching, N. (2023). Automated Annotation with Generative AI Requires Validation. *arXiv*, 2306.00176. https://doi.org/10.48550/arXiv.2306.00176

Pryzant, R., Iter, D., Li, J., Lee, Y.T., Zhu, C., & Zeng, M. (2023). Automatic Prompt Optimization with "Gradient Descent" and Beam Search. In H. Bouamor, J. Pino, & K. Bali (Eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing* (pp. 7957–7968). Association for Computational Linguistics. https://doi.org/10.48550/arXiv.2305.03495

Rathje, S., Mirea, D.M., Sucholutsky, I., Marjieh, R., Robertson, C.E., & Van Bavel, J.J. (2024). GPT Is an Effective Tool for Multilingual Psychological Text Analysis. *Proceedings of the National Academy of Sciences of the United States of America*, *121*(34), e2308950121. https://doi.org/10.1073/pnas.2308950121

Rogers, R. (2013). *Digital Methods*. Cambridge, MA: The MIT Press.

Sanh, V., Webson, A., Raffel, C., Bach, S., Sutawika, L., Zaid, A., Antoine, C., Arnaud, S., Arun, R., & Manan, D. (2022). Multitask Prompted Training Enables Zero-Shot Task Generalization. *arXiv*, 2110.08207. https://doi.org/10.48550/arXiv.2110.08207

Saravia, E. (2022). *Prompt Engineering Guide*. GitHub. https://github.com/dair-ai/Prompt-Engineering-Guide>

Sartor, G., & Lagioia, F. (2020). *The Impact of the General Data Protection Regulation (GDPR) on Artificial Intelligence*. Study. Panel for the Future of Science and Technology. EPRS, European Parliamentary Research Service. https://doi.org/10.2861/293

Sharma, S. (2019). *Data Privacy and GDPR Handbook*. Hoboken, NJ: Wiley. https://doi.org/10.1002/9781119594307

Spirling, A. (2023). Why Open-source Generative AI Models Are an Ethical Way Forward for Science. *Nature*, *616*, 413. https://doi.org/10.1038/d41586-023-01295-4

Tan, Z., Li, D., Wang, S., Beigi, A., Jiang, B., Bhattacharjee, A., Karami, M., Li, J., Cheng, L., & Liu, H. (2024). Large Language Models for Data Annotation: A Survey. *arXiv*, 2402.13446. https://doi.org/10.48550/arXiv.2402.13446

Törnberg, P. (2024a). *How to Use Large-Language Models for Text Analysis*. London: Sage. https://doi.org/10.4135/9781529683707

Törnberg, P. (2024b). Large Language Models Outperform Expert Coders and Supervised Classifiers at Annotating Political Social Media Messages. *Social Science Computer Review*, https://doi.org/10.1177/08944393241286471

Törnberg, P., & Uitermark, J. (2021). For a Heterodox Computational Social Science. *Big Data & Society*, *8*(2). https://doi.org/10.1177/20539517211047725

Weber, M., & Reichardt, M. (2023). Evaluation Is All You Need. Prompting Generative Large Language Models for Annotation Tasks in the Social Sciences. A Primer Using Open Models. *arXiv*, 2401.00284. https://doi.org/10.48550/arXiv.2401.00284

Wei, J., Bosma, M., Zhao, V.Y., Guu, K., Yu, A.W., Lester, B., Du, N., Dai, A.M., & Le, Q.V. (2022). Finetuned Language Models Are Zero-Shot Learners. *arXiv*, 2109.01652. https://doi.org/10.48550/arXiv.2109.01652

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q., & Zhou, D. (2024). Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, A. Oh (Eds.), *NIPS'22: Proceedings of the 36th International Conference on Neural Information Processing Systems* (pp. 24824–24837). New Orleans, LA: Curran.

White, J., Hays, S., Fu, Q., Spencer-Smith, J., Schmidt, D.C. (2024). ChatGPT Prompt Patterns for Improving Code Quality, Refactoring, Requirements Elicitation, and Software Design. In A. Nguyen-Duc, P. Abrahamsson, F. Khomh (Eds.), *Generative AI for Effective Software Development*. Cham: Springer. https://doi.org/10.1007/978-3-031-55642-5_4

Yan, C.T., Birks, M., & Francis, K. (2019). Grounded Theory Research: A Design Framework for Novice Researchers. *Sage Open Medicine*, *7*, 205031211882292. https://doi.org/10.1177/2050312118822927

Yu, H., Yang, Z., Pelrine, K., Godbout, J.F., & Rabbany, R. (2023). Open, Closed, or Small Language Models for Text Classification?. *arXiv*, 2308.10092. https://doi.org/10.48550/arXiv.2308.10092

Zhao, Z., Wallace, E., Feng, S., Klein, D., & Singh, S. (2021). Calibrate Before Use: Improving Few-Shot Performance of Language Models. In M. Meila & T. Zhang (Eds.), *Proceedings of the International Conference on Machine Learning* (pp. 12697–12706). https://doi.org/10.48550/arXiv.2102.09690

Ziegler, D.M., Stiennon, N., Wu, J., Brown, T.B., Radford, A., Amodei, D., Christiano, P., & Irving, G. (2019). Fine-Tuning Language Models from Human Preferences. *arXiv*, 1909.08593. https://doi.org/10.48550/arXiv.1909.08593

Zimmer, M. (2020). "But the Data Is Already Public": On the Ethics of Research in Facebook. In K.W. Miller & M. Taddeo (Eds.), *The Ethics of Information Technologies* (pp. 229–241). London: Routledge. https://doi.org/10.4324/9781003075011-17

Petter Törnberg – Institute for Language, Logic and Computation, University of Amsterdam (The Netherlands)
https://orcid.org/0000-0001-8722-8646 | p.tornberg@uva.nl
https://www.pettertornberg.com
Petter Törnberg is an Assistant Professor in Computational Social Science at the University of Amsterdam (The Netherlands). He studies the intersection of AI, social media, and politics, and draws on computational methods and digital data for critical inquiry. His recent books include *Intimate Communities of Hate: Why Social Media Fuels Far-Right Extremism* (with Anton Törnberg, Routledge, 2024), and *Seeing Like a Platform: An Inquiry into the Condition of Digital Modernity* (with Justus Uitermark; Routledge, forthcoming in early 2025).

Symposium: Repurposing Generative AI for Social Research – *peer-reviewed*
Edited by Federico Pilati, Anders Kristian Munk, and Tommaso Venturini
https://doi.org/10.6092/issn.1971-8853/19524

Sociologica. V.18 N.2 (2024)
ISSN 1971-8853
https://sociologica.unibo.it/

Integrating Large Language Models in Political Discourse Studies on Social Media: Challenges of Validating an LLMs-in-the-loop Pipeline

Giada Marino*[a] Fabio Giglietto[a]

[a] Department of Communication Sciences, Humanities and International Studies, University of Urbino Carlo Bo (Italy)

Submitted: May 11, 2024 – Revised version: August 20, 2024
Accepted: August 21, 2024 – Published: October 30, 2024

Abstract

The integration of Large Language Models (LLMs) into research workflows has the potential to transform the study of political content on social media. This essay discusses a validation protocol addressing three key aspects of LLM-integrated research: the versatility of LLMs as general-purpose models, the granularity and nuance in LLM-uncovered narratives, and the limitations of human assessment capabilities. The protocol includes phases for fine-tuning and validating a binary political classifier, evaluating cluster coherence, and assessing machine-generated cluster label accuracy. We applied this protocol to validate an LLMs-in-the-loop research pipeline designed to analyze political content on Facebook during the Italian general elections of 2018 and 2022. Our approach classifies political links, clusters them by similarity, and generates descriptive labels for clusters. This methodology presents unique validation challenges, prompting a reevaluation of accuracy assessment strategies. By sharing our experiences, this essay aims to guide social scientists in employing LLM-based methodologies, highlighting challenges and advancing recommendations for colleagues intending to integrate these tools for political content analysis on social media.

Keywords: Large Language Models (LLMs); Political Discourse; Social Media; Natural Language Processing (NLP).

Acknowledgements

This work was partially supported by the vera.ai project, which is funded by the European Union through the Horizon Europe program (Grant Agreement ID 101070093).

* giada.marino@uniurb.it

1 Introduction

Since ChatGPT's launch in November 2022, scholarly interest in Generative AI has grown significantly. A mini-review article published in August 2023 documented 156 Scopus-indexed publications referencing "ChatGPT" between November 2022 and April 2023 (Watters & Lemanski, 2023). As of April 2024, this number had surged to 4,642 publications for 2023 — with 1,303 in the social sciences — and 2,628 for 2024, with 622 in social sciences. This increase reflects widespread interest in generative AI's societal impacts and its integration into research practices, including those of social sciences (Rask & Shimizu, 2024), such as surveys, online experiments, and automated content analysis (Bail, 2024).

Large Language Models (LLMs), developed by organizations like OpenAI, Meta, Google, Anthropic, and Mistral AI, are versatile tools in natural language processing (NLP) workflows. These pre-trained models excel in general-purpose, prompt-based inferences and are widely used in chat-bot applications such as OpenAI's ChatGPT and Anthropic's Claude. Beyond chat-bots, LLMs' inferences are programmatically accessible and are known for their efficacy in zero-shot or few-shot learning tasks. They can be fine-tuned for specific needs across various domains. At their core, LLMs work by transforming text and multimedia content into numerical representations that capture core semantics. This process, referred to as embedding, is also performed by standalone embedding models and is currently used to enhance content retrieval in large datasets and support tasks like semantic search, clustering, topic modeling, and classification.

The potential of LLMs for text analysis and computational social science is widely recognized (Mu et al., 2024). However, concerns persist regarding their inherent limitations and biases (Grossmann et al., 2023), challenges with reproducibility (Balloccu et al., 2024; Chen et al., 2023), and the need for established best practices for their integration into research methodologies (Rask & Shimizu, 2024).

This essay contributes to the ongoing discourse by presenting a novel, fully LLM-integrated methodological pipeline, its text annotation, and analysis validation protocol. We focus on the unique challenges in validating such a pipeline, addressing a critical gap in current research on LLMs integration in social sciences. Our approach leverages state-of-the-art OpenAI models to uncover political narratives in Facebook-shared links during the 2018 and 2022 Italian general elections.

Our pipeline introduces LLMs in three ways: fine-tuning for binary classification of Italian political links, LLM-based embeddings for clustering similar political links, and direct API inferences for creating descriptive cluster labels.

Natural Language Processing (NLP) research has extensively employed transformer-based, fully fine-tuned models such as BERT, RoBERTa, DistilBERT, and XLNet to accomplish various tasks. However, despite the proliferation of domain-specific, language-specific, and task-specific versions, these models typically require fine-tuning before they can be effectively applied to specific tasks. Fine-tuning is both labor-intensive and computationally demanding (Bender et al., 2021). Once fine-tuned, the resulting model often performs well on the specific dataset, task, domain, or language, but its performance often degrades when any of these elements change. Traditional topic modeling algorithms, like Latent Dirichlet Allocation (LDA) and Latent Semantic Analysis (LSA), face significant challenges, including limitations related to the granularity of topics and issues (Abdelrazek et al., 2023). They also require a delicate and cumbersome preliminary text-cleaning phase and produce clusters of words that can often be difficult for researchers to interpret and utilize effectively (Gillings & Hardie, 2022).

Implementing a fully LLM-based pipeline presents significant validation challenges, necessitating a reevaluation of established accuracy assessment strategies. Drawing on the experience gained while designing and validating the pipeline, we explore the specific choices made during the validation protocol, focusing on three key characteristics of LLM-integrated research that complicate accuracy evaluation: the versatility of LLMs as general-purpose models, offering numerous application options with varying degrees of supervision, from multilingual capabilities, including underrepresented languages in research, to diverse content types, tasks, and fields of study; the varying levels of granularity and nuance in LLM-uncovered narratives; and the limitations of human assessment capabilities when evaluating models pre-trained on extensive datasets.

Our tailored validation protocol addresses these issues in three phases: fine-tuning and validating a binary political classifier, evaluating cluster coherence, and assessing the accuracy of machine-generated cluster labels. By sharing our experiences, this essay aims to provide insights for social scientists considering LLM-based research designs, highlighting both challenges and potential solutions in employing these advanced technologies in NLP.

2 Pipeline and Research Question

2.1 The Pipeline

Our LLMs-in-the-loop pipeline has five steps (Figure 1), including the identification of political links (2), embedding/clustering (3/4), and the generation of cluster labels (5). All these steps leverage the advanced capabilities of models provided by OpenAI.

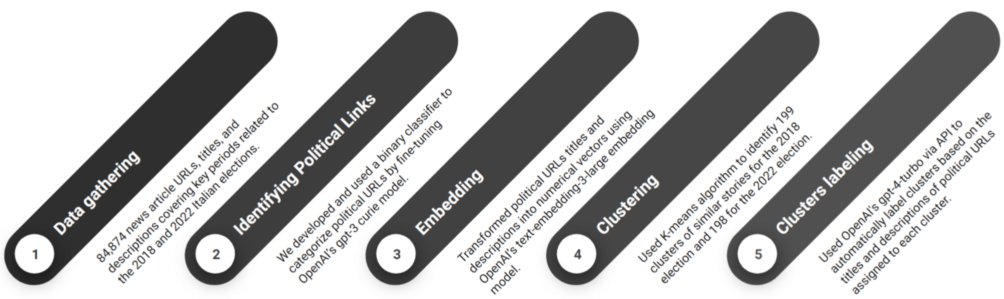

Figure 1: A graphical representation of the pipeline discussed in this article

2.1.1 Data Gathering

An initial dataset comprising 84,874 public news article URLs, titles, and descriptions was obtained by querying the Meta URL Shares Dataset for links first published on Facebook between December 24, 2017, and March 4, 2018, related to the 2018 election, and between July 21, 2022, and September 25, 2022, predominantly viewed by Italian users. Across the entire pipeline, only the title and description of these links (both are public content available at the respective URLs) have been used and thus fed to OpenAI's models.

2.1.2 Identifying Political Links

We developed a binary classifier to categorize political URLs by fine-tuning the GPT-3 *Curie* model, a now discontinued OpenAI model, suggested for this task. Seven Italian scholars with expertise in analyzing political news dissemination on social media supported the fine-tuning process. After standard training to ensure consistency (Krippendorff's alpha, Subjects = 200, Raters = 7, alpha = 0.812), they manually coded a proportional stratified (by-election and month) random sample of 4,190 URLs: 3,184 from 2018 and 1,006 from 2022. Excluding missing values and non-Italian URLs, the refined dataset for fine-tuning included 3,800 valid cases (1,801 political and 1,999 non-political). We concatenated titles and descriptions for each URL and filtered out non-Italian and empty titles and descriptions URLs, resulting in datasets of 59,838 URLs from 2018 and 17,690 from 2022. The classifier identified 54% of the 2018 posts and 53% of the 2022 posts as political, corresponding to 27,487 URLs in 2018 and 8,308 in 2022.

2.1.3 Grouping Together Similar Political Links

To identify clusters of similar links, we transformed our Italian political links into embeddings using a language model to convert the link's title and description into numerical vectors. After experimenting with various LLM-based embedding models (OpenAI's text-embedding-ada-002, Mistral AI's e5-mistral-7b-instruct (Wang et al., 2023), and OpenAI's text-embedding-3-large), we chose text-embedding-3-large based on clustering internal metrics. We preprocessed the text by removing HTML tags and hyperlinks before processing each URL's concatenated title and description.

Working with numerical vectors facilitates clustering-based topic modeling. Following OpenAI's recommendation, we used cosine distance to measure semantic similarity. We experimented with various clustering algorithms (k-means, DBSCAN, HDBSCAN, GenieClust (Gagolewski, 2021), and Kwikbucks (Silwal et al., 2023) and dimension reduction techniques (t-SNE and UMAP). Ultimately, we implemented cluster analysis using k-means with Lloyd's algorithm and retained all the initial 3,072 dimensions. Moreover, Giglietto's (2024) research on the same dataset demonstrated that LLMs outperform fully fine-tuned transformer models in NLP tasks. To determine the optimal number of clusters, we employed Bayesian optimization aimed at maximizing the Silhouette score and Hplus metric Dyjack et al., 2023), ranging from 2 to 200 clusters, with 200 considered the maximum number for interpretability at the level of granularity requested by the scope of the study. This process identified 199 clusters as optimal for the 2018 election and 198 clusters for the 2022 election.

2.1.4 Clusters Labeling

We used GPT-4-turbo through the API to programmatically label clusters based on their content. The process involves feeding the model a sample of items from each cluster and requesting a short descriptive label. Table A3 (Appendix A) reports on the prompt specifics. The prompt includes a system prompt for context and output format and a user prompt supplying necessary documents.

The final prompt was crafted using strategies from OpenAI's prompt engineering guide (OpenAI, 2024) and tested for consistency across multiple runs and GPT models. The process of optimization was mainly aimed at instructing the model to output the specific label, only avoiding any further premise (e.g., "The label of the cluster is..") or comment. We prioritized a detailed prompt over cost optimization. Costs are computed per token, with different values for input and output tokens.

The model employed to generate the cluster labels includes training data up to December 2023, encompassing the 2022 election period. The total cost to label 199 clusters from 2018 and 198 from 2022 was $30. Each label was requested using a prompt combining standard text with a density-based sample of cluster items. Despite GPT-4-turbo's 128,000 token capacity, we limited prompts to 8,000 tokens for a fair comparison with GPT-4. This approach achieved an average coverage of 84% of items per cluster.

2.2 Research Question

This essay focuses on how a full LLM-supported pipeline for social media political content annotation can be validated. This methodological approach poses several challenges, including issues with model reliability, data interpretation, integration of these models into existing research frameworks, and the relative newness of studies relying on these tools. For these reasons, this essay seeks to answer the following research question:

What are the main challenges researchers may face in validating an LLM-in-the-loop methodological approach, and how can they be addressed?

To answer this research question, each of the following sections of this essay is dedicated to a specific challenge we faced during our introduction of LLMs for text annotation tasks.

The next section discusses the general-purpose nature of LLMs, which are pre-trained on large datasets and perform general-purpose tasks based on given prompts. This means they can understand and generate text based on various inputs and handle different kinds of tasks, making them useful in many applications. Considering this adaptability, they may be employed, supervised or not, at different steps of a multiple research pipeline. This characteristic requires novel, tailored validation approaches. To address this, our protocol comprises three distinct phases, each corresponding to different LLM applications in our study.

The fourth section tackles the theoretical challenge of narrative definition. The possibility of unsupervised generation of embeddings with an almost unlimited number of dimensions and clustering them results in a cluster granularity that necessarily affects the validation preparation phase. This granularity ranges from general topics to specific journalistic stories. Our approach utilizes multi-level, detailed validation guidelines to ensure the accuracy and relevance of our findings.

The fifth and final section addresses the "knowledge" challenge. LLMs, having been trained on broad datasets, possess competencies that often surpass those of traditional coders. This necessitates a careful selection of the number and profiles of the coder team to ensure that the validation process is both thorough and effective.

3 Tailoring Validation Protocols for General-Purpose LLMs in Text Classification Tasks

While LLMs can be applied to a wide range of NLP tasks, their very complexity also means that their use in specific research contexts requires a significant degree of human guidance and decision-making. Unlike more narrowly defined machine learning models, LLMs are primarily designed for general-purpose, prompt-based inferences (Huang et al., 2023; Kuzman et al., 2023). Adapting them to meet the needs of a particular research objective or workflow involves a number of crucial human-led decisions at multiple stages.

In our case study, we needed to select a specific embedding model and clustering algorithm, as well as determine clustering parameters such as the number of clusters, the labeling model and prompt, and the sample size to input into the model. All these decisions require justification and must be evaluated against alternative options. Training a classifier, evaluating the performance of unsupervised cluster analysis, and extracting cluster narratives' require a training set or a ground truth to assess the LLMs' effectiveness in these tasks.

Researchers rely on different strategies and techniques to evaluate model fit. The prevailing methodologies for assessing the performance of LLMs typically involve a range of standardized tests covering areas from common sense reasoning and reading comprehension to arithmetic and coding. Although extensive, these benchmarks often fail to fully explore the nuanced capabilities afforded by a natural language interface. For instance, while they measure accuracy in specific tasks, they may not adequately assess the model's ability to handle ambiguous or contextually rich scenarios, nor do they always test for biases or the generation of novel content.

Similarly, embedding models are evaluated across diverse tasks such as classification, clustering, retrieval, and summarization (Muennighoff et al., 2022). However, these evaluations generally focus on optimizing straightforward metrics like accuracy or F1 scores, which might not capture more subjective qualities like the relevance or coherence of the content generated. Moreover, the language dependency of these tests presents another significant limitation. Most benchmarks are developed in English and subsequently translated for other languages, potentially skewing performance assessments due to translation inaccuracies or cultural nuances not being adequately represented. This approach can obscure the true versatility and effectiveness of LLMs and embedding models in non-English contexts, hence limiting our understanding of their global applicability and efficacy.

Given the limitations of automatic validation methods (Clinciu et al., 2021; Iskender et al., 2020), human evaluation has increasingly been recognized as a critical component in NLP research, either complementing or replacing these methods (Schuff et al., 2023). Validation protocols involving human teams require them to address specific research questions by following detailed guidelines, particularly when researchers test precise hypotheses (Schuff et al., 2023).

Given the general purpose nature of LLMs and the characteristics of our dataset, which includes a specific social media platform (Facebook), the domain (politics), and language-specific elements (Italian), the three-way LLMs are implemented in our workflow necessitated a distinct and tailored validation protocol to assess its efficacy. This implies that validating different LLM applications through existing standard processes employed for fully fine-tuned transformer models is challenging, and specific validation protocols are still under development. Furthermore, a validation workflow customized for our study might not be universally applicable or extendable to other datasets or domains.

In light of these considerations, we developed a three-step, ad hoc validation protocol. We opted for evaluation protocols involving human annotators.

The first round of validation pertained to the binary classifier of political vs. non-political URLs and was conducted employing standard validation approaches and measures. The fine-tuning dataset, manually labeled by seven human experts, was divided into training and validation sets. The training set was used to fine-tune the model, while the validation set assessed its performance, achieving an F1 score of 0.897, with a precision of 0.911 and a recall of 0.883.

The second round of validation regards a different task we accomplished through the use of LLMs, specifically cluster analysis. This phase involves assessing the coherence of clusters. Six expert coders, familiar with political content on social media and the Italian political landscape, evaluated a sample selected through systematic sequential pairing followed by a random subsampling, which comprised either 10% or at least five pairs from each cluster, totaling 2,754 pairs for the 2018 elections and 994 pairs for the 2022 elections. Coders were presented with pairs of links (Grimmer & King, 2011) from the same cluster and were required to assign a coherence level based on guidelines established during preliminary training.

Following the preliminary training, the coders were divided into three teams of two, each team comprising one experienced and one less experienced coder. Each team was assigned to a random subset of one-third of the items in the evaluation sample. Both team members independently coded the assigned pairs and held two meetings — a preliminary alignment meeting and a concluding meeting — to resolve any discrepancies in their evaluations with their teammates.

The guidelines provided to the coders (see Table A1 in Appendix A) use a scale ranging from 0, indicating a lack of coherence, to 4, indicating two items belonging to the same journalistic story, with an additional level for non-codable pair cases.

The last round of validation concerned the machine-generated labels. The goal is to evaluate how accurately each label represents the content it is intended to describe. The evaluation was carried out by the same six coders involved in the evaluation of the clusters' coherence. Following a phase of training performed on a pilot subsample of one item (and its respective label) for each cluster (199 for 2018 and 198 for 2022), the team agreed on a codebook consisting of four criteria (thematic alignment, implications, content coverage, and contextual alignment) and a three-level scale (misfit, partial fit, and good fit). Each coder is asked to rate the accuracy of a cluster label for one of the items assigned to that cluster. The evaluation employs a density-based sampling approach where each cluster contributes either a minimum of 10 items or 10% of its total, whichever is greater. This method ensures that each cluster is adequately represented in the sample. Specifically, the sampling technique is designed to represent proportionally the variety of centroid distances within each cluster. Focusing on density rather than a uniform distribution, the method ensures coverage across all regions of the distance distribution, from the closest to the furthest items from the centroid.

4 Validating LLMs-detected Political Narrative: Addressing Challenges in Theoretical Definition

Researchers have utilized various automatic classification methods to group similar social media political content and label them (Gupta et al., 2020). With the rise of social media as one of the primary news sources, narrative detection has become increasingly relevant. This is partly due to algorithmic indexing, which amplifies content based on its popularity, allowing certain narratives to gain more attention and thus be shown to more users.

However, defining the specific conditions that qualify a sequence of words or sentences as a narrative remains contentious in content annotation research. Despite its relevance, scholars have struggled to reach a consensus on a definition. Generally, scholars agree that "narrative is a key concept for understanding human behavior and beliefs" (Piper et al., 2021, p. 298). Consistency in terminology is crucial for clearly defining the boundaries of the research object and setting the study's objectives, particularly in NLP research, where a precise interpretation of linguistic phenomena is required.

In narratology, "narrative" refers to the structure of events involving a complex set of features such as time, context, participants, and the narrator's perspective in organizing information (Genette, 1980; Pianzola, 2018; Piper et al., 2021).

In the analysis of political discourses, terms like "topics" or "issues" are more frequently used within the theoretical framework of the public agenda (Boydstun, 2013). These terms serve as cognitive shortcuts that describe aspects of reality and vary in attention based on media coverage, thereby influencing public debate and political decisions (Scheufele, 2000).

A "political issue" is a subcategory of a topic describing an event or a series of events perceived as a significant problem by citizens (Wlezien, 2005). In political communication, various institutions and researchers label groups of content to study, for example, the main topics or issues of political parties and candidates' campaigns (Illuminating, 2020) or disinformation during the elections in several European countries (EDMO, 2024).

The term "narrative" is less utilized in political communication studies because it is often conflated with storytelling or used in other scientific areas, such as linguistics. Groth (2019) refers to Eagleton (1979), who argued that narratives present closed stories with coherent logic, offering stringent explanations, causal relationships, and genealogies for socio-cultural and political realities. In this view, it is close to the definition of the more commonly used concept of media frames (Matthes & Kohring, 2008; McCombes et al., 2006; Reese, 2007)]. Also, Bradshaw et al. (2024) provide an insightful framework for examining "strategic narratives" in Russian discourse about the Ukrainian conflict. Drawing on prior literature, Popkova (2023) and Schmitt (2018) identify three key types of narrative manifestations: narratives related to international relations and global "world order," identity narratives tied to a country's culture and traditions, and issue-specific narratives focused on particular topics.

Also, Kotseva et al. (2023), employ a multidimensional hierarchical definition of narrative ranging from sub-narrative to super-narrative. Particularly interesting is the super-narrative definition. In comparison with the narrative, the super-narrative has a cross-temporal and cross-country nature, as a story-line that survives and evolves over time takes advantage every time and in different contexts of single events or local specificities.

In this fragmented scenario, the boundaries of a narrative are left to the discretion of researchers. When using supervised or semi-supervised methods, a tailored narrative definition is essential when setting a codebook for fine-tuning a transformer-based model for content annotation or cleaning datasets to achieve refined results (Groth, 2019; Kotseva et al., 2023). These approaches require the researcher to clearly delineate the scope and} characteristics of the narratives upfront. In contrast, when using topic modeling techniques such as LDA, the dimensions of a narrative are left more open to interpretation based on the analysis outcomes.

Approaches that leverage LLMs for unsupervised or minimally supervised content annotation can produce results with varying levels of detail and granularity. As discussed earlier, this variability is not necessarily a weakness but rather a strength that allows for more nuanced and contextual findings.

In our own analysis, we took a theoretical holistic view, considering narratives as common story-lines that tap into collective memories, emotions, and historical analogies to achieve political objectives — aligning with the broader vision outlined by Bradshaw et al. (2024).

Challenges arise when researchers must validate these clustering outcomes. This process necessitates adaptable validation protocols that can assess different levels of coherence and accuracy. Clusters identified by k-means algorithms for our case tend to vary in both size and specificity. Some clusters are more generic, encompassing a range of closely related issues, while others are highly specific, tied to a single journalistic story or a particular media frame. This diversity in cluster characteristics underscores the need for flexible and robust validation methods.

To mitigate this issue during validation, we implemented some adaptation actions. Firstly, we evaluated cluster coherence by rating the coherence to random pairs of links extracted from each cluster (the guidelines are detailed in Table A1). We split the evaluation of coherence into three distinct levels. The basic level of coherence pertains to the topic as a broad area belonging to politics, such as economy, health, immigration, environment, safety, etc. A second, more specific level of coherence refers to stories with the same actor, event, place, or organization in common. Level three is the narrowest coherence estimation, and it regards only those pairs that refer to the same journalistic story, e.g., the murder case of Pamela Mastropietro in 2018. We also added a level 98 to indicate ambiguous cases or when the coder is uncertain. At the end of the coding phase, teammates discussed these specific cases to assign them another value in the scale.

We utilized a scale specifically designed to validate the accuracy levels of cluster labels generated by GPT-4-turbo. In contrast to assessing the coherence of the cluster — which relies on an established algorithm and innovative embeddings derived from Italian text — the application of an LLM to label the clusters is less conventional.

Moreover, these labels are critical for the subsequent phase of our research design, where exposure and engagement metrics will be calculated and analyzed based on the labels' meanings. Therefore, accurately assessing the labels' ability to represent the underlying content of each cluster is essential.

It is important to note that the two validation processes, though aimed at distinct tasks, are interconnected. A lack of coherence within a cluster would indeed hinder the creation of meaningful and representative short labels.

The rating scale adopted for evaluating the labels ranges from one (Misfit) to three (Good fit) (see Table A2 in Appendix A). The evaluation of label fit is based on four criteria established by the team of coders during the alignment meeting. More specifically, these criteria include:

- The thematic alignment criterion measures the extent to which the label corresponds to the central themes or subjects discussed in the item. Thematic alignment verifies that the label directly includes the primary topic addressed by the item.
- The implications or connotations suggested by the label. This criterion checks whether the label implies any outcomes, consequences, or broader trends consistent with the information or narrative provided in the item, ensuring that the label does not exaggerate, oversimplify, or misrepresent the content's potential impacts or significance.
- The content coverage standard assesses if the label encapsulates the key elements, facts, and details presented in the item. Content coverage ensures that the label addresses all significant points, leaving no major aspect of the content unrepresented or inaccurately portrayed. Additionally, a label should not encompass themes or details that extend beyond the scope of the item, which could mislead the understanding of the item's focus.
- The contextual alignment criterion evaluates the label's accuracy in reflecting the item's geographical, cultural, historical, or situational context. Contextual alignment confirms

that the label is suitable for the specific setting in which the content is placed, adhering to any particular nuances that influence content understanding.

The lowest value is attributed when a label completely fails to align with the item's content. The partial fit judgment is assigned when the label relates to the item in terms of theme and implications, but either the content covered by the label is too narrow or broad, or its context diverges from the item's context. This is the case, for example, with news discussing the rise in unemployment rates, specifically in rural parts of Italy due to local factory closures, and the label generated by the LLM for the cluster is "*Economic Challenges in the European Union*." There is a good fit between a label and a piece of news when it accurately represents the item across all aspects. We consider a good fit case, for example, news related to rescue operations off the Sicilian coast highlighting the ongoing challenges faced by migrants and labeled as "*Migrant Crisis and Humanitarian Efforts in the Mediterranean*."

5 Dealing with LLMs' High Levels of Knowledge

Assessing the reliability of LLM content annotation is a fundamental step in the process (Chiang & Lee, 2023; Gilardi et al., 2023), particularly challenging within complex research designs. Despite the enormous analytical opportunities and creative potential afforded by LLMs (Gilardi et al., 2023; Jahan et al., 2023), human evaluation remains essential.

Historically, human evaluation has been crucial to understanding the performance of natural language processing (NLP) models or algorithms (Gillick & Liu, 2010; Guzmán et al., 2015). We rely on human evaluators because certain textual aspects are difficult to assess with automatic evaluation metrics, necessitating human judgment either to train the model or to rate the quality of its outputs. However, human evaluation is known for its instability (Clark et al., 2021; Gillick & Liu, 2010), attributed to factors ranging from the quality of the workforce (Karpinska et al., 2021) to challenges in reproducing the same tasks or training human experts to provide consistent assessments (Chiang & Lee, 2023). Despite these limitations, human evaluation is prevalent and commonly considered indispensable in NLP, offering advantages over automatic metrics when carefully implemented.

In addition to the training phases, the most relevant task in models of human validation is recruiting the most appropriate team of annotators for the task. Primary strategies for recruiting annotators include hiring and training coders, such as students or research assistants, or utilizing crowdsourced work services like Amazon Mechanical Turk (MTurk) (Kasthuriarachchy et al., 2021). These strategies may be used individually or in combination, with trained coders annotating relatively small datasets considered gold standards and crowd workers increasing the volume of annotations (Gilardi et al., 2023). However, the limitations of these approaches increase when using LLMs for content annotation tasks, as they have been shown to outperform crowd workers, especially in complex tasks (Gilardi et al., 2023; Huang et al., 2023; Törnberg, 2023). Specifically, in the context of political communication research, LLMs possess significant knowledge of political and cultural contexts compared to a low-skilled workforce. Additionally, recruiting students or research assistants with deep knowledge of the political context is challenging, and this specific training is extremely time-consuming and resource-intensive. Furthermore, when conducting research in less commonly spoken languages, such as those other than English or Spanish, the recruitment process becomes complicated due to the scarcity of native-speaking crowd workers.

We conducted our validation rounds with these challenges in mind. Initially, we considered recruiting crowd workers from Fiverr, a platform that facilitates the hiring of Italian freelancers, and selected eight coders with expertise in copy-editing and data analysis.

However, after careful consideration, we decided against using crowd workers for validating our results. During the validation phase of our pipeline, we needed to thoughtfully select evaluators to assess the quality of clustering and labeling. As previously discussed, studies have shown that crowd workers may underperform compared to large language models in certain content annotation tasks (Gilardi et al., 2023; Huang et al., 2023; Törnberg, 2023; Zhang et al., 2023). This prompted us to reconsider whether crowd workers would be the most appropriate judges for an approach that surpasses their performance in the same tasks. For instance, when assessing the accuracy of the clustering, we encountered several cases that were challenging even for experts familiar with the national political context. This difficulty arises because it is unreasonable to expect humans to recall every specific political event and actor over the years. To illustrate with an example from our dataset, during one of the coder training sessions in the validation phase, we encountered the following story included in the cluster labeled as "*Corruption and Criminal Allegations in Italian Politics and Public Services*":

> Amedeo Matacena has died: struck down by a sudden illness. Matacena died at 59 years old in Abu Dhabi, where he had been living for years. The former Forza Italia deputy Amedeo Matacena, a well-known entrepreneur from Reggio Calabria, died at the age of 59. He was the son of the shipowner of the same name who passed away in 2003, and he was famous for initiating the ferry service across the Strait of Messina with Caronte [...] (translated from the original in Italian).

At a first look, this news seems to deal with the death of a secondary, former Italian politician. It was necessary to google the name Amedeo Matacena to discover that he had been convicted of involvement in a mafia association and had been a fugitive in Abu Dhabi until his death.

Given these challenges, we thus decided to rely on expert researchers in political communication to conduct the three validation rounds. As mentioned in the previous paragraph, we employed a team of seven coders for the fine-tuning and validation phase of the binary political classifier. This team consisted of all the authors of a paper we presented at the annual conference of the Italian Political Communication Association in 2023. Except for one PhD student, all co-authors are postdoctoral researchers and associate professors specializing in political communication and social media studies, and all are native Italian speakers.

In the second and third rounds of validation, we employed a team of six expert coders, four of whom had also participated in the first round. In this instance, the annotators were all PhD candidates, postdoctoral researchers, and associate professors focusing on political communication and social media research topics and all were native or proficient in Italian.

The less expert researchers were trailed and supervised by the more proficient ones, in particular concerning knowledge of the last ten years' Italian political scenario. The processes of the second and third rounds of validation are described extensively in Section Three.

6 Conclusions

In this work, we pioneer the exploration of multiple validation protocols for different tasks in political discourse annotation using LLMs. Incorporating LLMs in natural language processing marks a significant paradigm shift within the field, offering a viable and adaptable method for mostly unsupervised clustering analysis and narrative extraction. Indeed, they bring the potential of transformer language models like BERT to topic modeling methods (Mu et al., 2024). LLMs demonstrate their capability to handle specific domain, platform, and cultural context datasets with little to no fine-tuning required.

Thanks to their general-purpose nature, LLMs can manage extensive and complex tasks, enabling elaborate methodological pipelines. Specifically, we employed LLMs for three different tasks on two datasets of Facebook links related to the 2018 and 2022 Italian elections. We thus used LLMs in model fine-tuning to build a highly reliable binary classifier of political and non-political links, to generate LLM-based embeddings to cluster similar political content, and to make inferences via API to create short descriptive labels for the identified clusters.

However, using LLMs in all the steps of our NLP pipeline also introduces several new challenges, particularly in validating methodologies. We faced major challenges, particularly when we evaluated the outcomes of the unsupervised tasks, such as cluster analysis and label generation. At a general level, an LLM-in-the-loop pipeline necessitates distinct and tailored validation steps to assess the efficacy of each of the pipeline actions. In cluster analysis outcomes, we observed that LLMs can generate clusters/narratives with varying granularity levels, affecting how we consider the items within the same narrative group accurate or coherent and requiring highly detailed and adaptable codebooks. Moreover, LLMs' deep understanding of political and cultural contexts impacts the selection of the workforce for validation processes involving human participants, challenging traditional methods of recruiting content annotators and making the involvement of high-profile experts necessary. The versatility of LLMs encourages the phasing out of outdated annotation methods previously used in NLP studies. For instance, reliance on a low-skilled workforce annotation through crowdsourcing services like Amazon Mechanical Turk may become less necessary, as LLMs can efficiently process and understand large datasets with greater accuracy. This shift necessitates the development of new, robust validation protocols that keep pace with the rapid advancements in machine learning and artificial intelligence. These protocols must ensure that the models are not only effective but also free from bias and ethically compliant. Our validation protocol, for example, attempts to address potential biases by implementing a human-led task-by-task evaluation that relies fully on experts (Pangakis et al., 2023). Regarding the ethical concern of using models from proprietary providers for political content annotation, we mitigated this issue by choosing to provide the model with titles and brief descriptions of news stories that are already publicly available. Thus, we did not expose any proprietary, private, or sensitive information to the model.

The development of validation protocols for using LLMs to analyze the digital political discourse is a compelling issue. A timely implementation of LLMs in this field of studies may, in fact, be crucial in preventing their misuse.

Over the last two decades, each technological tool producing information flows has been susceptible to exploitation by malicious actors to spread problematic information and manipulate public opinion. In this context, LLMs can act as a double-edged sword. Prompt and competent adoption of these tools by political communication and science researchers may be pivotal in preventing or tackling such abuses and safeguarding the integrity of information while promoting responsible technology use in society.

Overall, the advancement of LLMs in NLP has opened new avenues for research and application. Although our research design is particularly complex, including various rounds of annotation that exploited LLMs for different tasks, the resources consumed in terms of time, costs, and researchers involved are, considering the scale of the project, limited.

Finding effective validation solutions that minimize the challenges of implementing an

LLMs-in-the-loop pipeline for content annotation may facilitate the introduction of LLMs into social science research. We wrote this essay to share our experience and expect it to serve as a guide to other researchers who would introduce LLMs in their studies. We hope that sharing our knowledge can contribute to the early adoption of similar methodological approaches using LLMs for digital political content annotation.

References

Abdelrazek, A., Eid, Y., Gawish, E., Medhat, W., & Hassan, A. (2023). Topic Modeling Algorithms and Applications: A Survey. *Information Systems*, *112*, 1–17. https://doi.org/10.1016/j.is.2022.102131

Bail, C.A. (2024). Can Generative AI Improve Social Science? *Proceedings of the National Academy of Sciences of the United States of America*, *121*(21), 1–10. https://doi.org/10.1073/pnas.2314021121

Balloccu, S., Schmidtová, P., Lango, M., & Dušek, O. (2024). Leak, Cheat, Repeat: Data Contamination and Evaluation Malpractices in Closed-Source LLMs. *arXiv*. https://doi.org/10.48550/arXiv.2402.03927

Bender, E.M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (pp. 610–623). Association for Computing Machinery. https://doi.org/10.1145/3442188.3445922

Boydstun, A.E. (2013). *Making the News: Politics, the Media & Agenda Setting*. Chicago, IL: University of Chicago Press.

Bradshaw, S., Elswah, M., Haque, M., & Quelle, D. (2024). Strategic Storytelling: Russian State-Backed Media Coverage of the Ukraine War. *International Journal of Public Opinion Research*, *36*(3), edae028. https://doi.org/10.1093/ijpor/edae028

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

Chiang, C.-H., & Lee, H.-Y. (2023). Can Large Language Models Be an Alternative to Human Evaluations? *arXiv*. https://doi.org/10.48550/arXiv.2305.01937

Clark, E., August, T., Serrano, S., Haduong, N., Gururangan, S., & Smith, N. A. (2021). All That's "Human" Is Not Gold: Evaluating Human Evaluation of Generated Text. *arXiv*. https://doi.org/10.48550/arXiv.2107.00061

Clinciu, M., Eshghi, A., & Hastie, H. (2021). A Study of Automatic Metrics for the Evaluation of Natural Language Explanations. *arXiv*. https://doi.org/10.48550/arXiv.2103.08545

Dyjack, N., Baker, D.N., Braverman, V., Langmead, B., & Hicks, S.C. (2023). A Scalable and Unbiased Discordance Metric with H. *Biostatistics*, *25*(1), 188–202. https://doi.org/10.1093/biostatistics/kxac035

Eagleton, T. (1979). Ideology, Fiction, Narrative. *Social Text*, *2*, 62–80. https://doi.org/10.2307/466398

European Digital Media Observatory (EDMO). (2024). *Disinformation Narratives during the 2023 Elections in Europe*. https://edmo.eu/publications/second-edition-march-2024-disinformation-narratives-during-the-2023-elections-in-europe/

Gagolewski, M. (2021). genieclust: Fast and Robust Hierarchical Clustering. *SoftwareX*, *15*, 100722. https://doi.org/10.1016/j.softx.2021.100722

Genette, G. (1980). *Narrative Discourse: An Essay in Method*. (J.E. Lewin, Trans.). Ithaca, NY: Cornell University Press. (Original work published 1972)

Giglietto, F. (2024). Evaluating Embedding Models for Clustering Italian Political News: A Comparative Study of Text-Embedding-3-Large and UmBERTo. *OSF Preprints*. https://doi.org/10.31219/osf.io/2j9ed

Gilardi, F., Alizadeh, M., & Kubli, M. (2023). ChatGPT Outperforms Crowd Workers for Text-annotation Tasks. *Proceedings of the National Academy of Sciences of the United States of America*, *120*(30), e2305016120. http://doi.org/10.1073/pnas.2305016120

Gillick, D., & Liu, Y. (2010). Non-expert Evaluation of Summarization Systems is Risky. In C. Callison-Burch, & M. Dredze (Eds.), *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk* (pp. 148–151). Association for Computational Linguistics. https://aclanthology.org/W10-0722

Gillings, M., & Hardie, A. (2022). The Interpretation of Topic Models for Scholarly Analysis: An Evaluation and Critique of Current Practice. *Digital Scholarship in the Humanities*, *38*(2), 530–543. https://doi.org/10.1093/llc/fqac075

Grimmer, J., & King, G. (2011). General Purpose Computer-assisted Clustering and Conceptualization. *PNAS*, *108*(7), 2643–2650. https://doi.org/10.1073/pnas.1018067108

Grossmann, I., Feinberg, M., Parker, D.C., Christakis, N.A., Tetlock, P.E., & Cunningham, W.A. (2023). AI and the Transformation of Social Science Research. *Science*, *380*(6650), 1108–1109. https://doi.org/10.1126/science.adi1778

Groth, S. (2019). Political Narratives / Narrations of the Political: An Introduction. *Narrative Culture*, *6*(1), 1–18. https://doi.org/10.13110/narrcult.6.1.0001

Gupta, S., Bolden, S., & Kachhadia, J. (2020). *PoliBERT: Classifying Political Social Media Messages with BERT* (Working paper SBP-BRIMS 2020 conference). Social, Cultural. https://news.illuminating.ischool.syr.edu/2020/11/24/polibert-classifying-political-social-media

Guzmán, F., Abdelali, A., Temnikova, I., Sajjad, H., & Vogel, S. (2015). How Do Humans Evaluate Machine Translation. In O. Bojar, R. Chatterjee, C. Federmann, B. Haddow, C. Hokamp, M. Huck, V. Logacheva, & P. Pecina (Eds.), *Proceedings of the Tenth Workshop on Statistical Machine Translation* (pp. 457–466). Association for Computational Linguistics. https://aclanthology.org/W15-3059/

Huang, F., Kwak, H., & An, J. (2023). Is ChatGPT Better than Human Annotators? Potential and Limitations of ChatGPT in Explaining Implicit Hate Speech. In Y. Ding, J. Tang, J. Sequeda, L. Aroyo, C. Castillo, & G.-J. Houben (Eds.), *Companion Proceedings of the ACM Web Conference 2023* (pp. 294–297). ACM Digital Library. https://doi.org/10.1145/3543873.3587368

Illuminating. (2020). *2020 Presidential Campaign Facebook and Instagram Ads*. https://illuminating.ischool.syr.edu/campaign_2020/

Iskender, N., Polzehl, T., & Möller, S. (2020). Best Practices for Crowd-based Evaluation of German Summarization: Comparing Crowd, Expert and Automatic Evaluation. In S. Eger, Y. Gao, M. Peyrard, W. Zhao, & E. Hovy (Eds.), *Proceedings of the First Workshop on Evaluation and Comparison of NLP Systems* (pp. 164–175). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.eval4nlp-1.16

Jahan, I., Laskar, M.T.R., Peng, C., & Huang, J. (2023). Evaluation of ChatGPT on Biomedical Tasks: A Zero-Shot Comparison with Fine-Tuned Generative Transformers. *arXiv*. https://doi.org/10.48550/arXiv.2306.04504

Karpinska, M., Akoury, N., & Iyyer, M. (2021). The Perils of Using Mechanical Turk to Evaluate Open-ended Text Generation. In M.F. Moens, X. Huang, L. Specia, & S. Wen-tau Yih (Eds.), *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing* (pp. 1265–1285). Association for Computational Linguistics. https://10.18653/v1/2021.emnlp-main.97

Kasthuriarachchy, B., Chetty, M., Shatte, A., & Walls, D. (2021). Cost Effective Annotation Framework Using Zero-shot Text Classification. *Proceedings of the 2021 International Joint Conference on Neural Networks* (pp. 1–8). IEEE. https://doi.org/10.1109/IJCNN52387.2021.9534335

Kotseva, B., Vianini, I., Nikolaidis, N., Faggiani, N., Potapova, K., Gasparro, C., Steiner, Y., Scornavacche, J., Jacquet, G., Dragu, V., Della Rocca, L., Bucci, S., Podavini, A., Verile, M., Macmillan, C., & Linge, J. (2023). Trend Analysis of COVID-19 Mis/Disinformation Narratives: A 3-year Study. *PLOS ONE*, *18*(11), 1–26. https://doi.org/10.1371/journal.pone.0291423

Kuzman, T., Mozetic, I., & Ljubešic, N. (2023). ChatGPT: Beginning of an End of Manual Linguistic Data Annotation. Use Case of Automatic Genre Identification. *arXiv*. https://doi.org/10.48550/arXiv.2303.03953

Matthes, J., & Kohring, M. (2008). The Content Analysis of Media Frames: Toward Improving Reliability and Validity. *The Journal of Communication*, *58*(2), 258–279. https://doi.org/10.1111/j.1460-2466.2008.00384.x

McCombes, M., Lopez-Escobar, E., & Llamas, J.P. (2006). Setting the Agenda of Attributes in the 1996 Spanish General Election. *The Journal of Communication*, *50*(2), 77–92. https://doi.org/10.1111/j.1460-2466.2000.tb02842.x

Muennighoff, N., Tazi, N., Magne, L., & Reimers, N. (2022). MTEB: Massive Text Embedding Benchmark. *arXiv*. https://doi.org/10.48550/ARXIV.2210.07316

Mu, Y., Dong, C., Bontcheva, K., & Song, X. (2024). Large Language Models Offer an Alternative to the Traditional Approach of Topic Modelling. *arXiv*. http://arxiv.org/abs/2403.16248

OpenAI. (2024). *Prompt Engineering*. https://platform.openai.com/docs/guides/prompt-engineering/prompt-engineering

Pangakis, N., Wolken, S., & Fasching, N. (2023). Automated Annotation with Generative AI Requires Validation. *arXiv*. https://doi.org/10.48550/arXiv.2306.00176

Pianzola, F. (2018). Looking at Narrative as a Complex System: The Proteus Principle. In R. Walsh & S. Stepney (Eds.), *Narrating Complexity* (pp. 101–122). NY, New York: Springer International Publishing.

Piper, A., So, R.J., & Bamman, D. (2021). Narrative Theory for Computational Narrative Understanding. In Moens, M.-F., Huang, X., Specia, L., & Yih, S. W.-T. (Eds.), *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing* (pp. 298–311). Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.emnlp-main.26

Popkova, A. (2023). Strategic Narratives of Russiagate on Russian Mainstream and Alternative Television. In O. Boyd-Barrett & S. Marmura (Eds.), *Russiagate Revisited: The Aftermath of a Hoax* (pp. 203–223). NY, New York: Springer International Publishing.

Rask, M., & Shimizu, K. (2024). Beyond the Average: Exploring the Potential and Challenges of Large Language Models in Social Science Research. *Proceedings of the 2024 International Conference on Artificial Intelligence, Computer, Data Sciences and Applications* (pp. 1–5). IEEE. https://doi.org/10.1109/ACDSA59508.2024.10467341

Reese, S.D. (2007). The Framing Project: A Bridging Model for Media Research Revisited. *The Journal of Communication*, *57*(1), 148–154. https://doi.org/10.1111/j.1460-2466.2006.00334.x

Scheufele, D.A. (2000). Agenda-Setting, Priming, and Framing Revisited: Another Look at Cognitive Effects of Political Communication. *Mass Communication and Society*, *3*(2–3), 297–316. https://doi.org/10.1207/S15327825MCS0323_07

Schmitt, O. (2018). When Are Strategic Narratives Effective? The Shaping of Political Discourse through the Interaction between Political Myths and Strategic Narratives. *Contemporary Security Policy*, *39*(4), 487–511. https://doi.org/10.1080/13523260.2018.1448925

Schuff, H., Vanderlyn, L., Adel, H., & Vu, N.T. (2023). How to Do Human Evaluation: A Brief Introduction to User Studies in NLP. *Natural Language Engineering*, *29*(5), 1199–1222. https://doi.org/10.1017/S1351324922000535

Silwal, S., Ahmadian, S., Nystrom, A., McCallum, A., Ramachandran, D., & Kazemi, S.M. (2023). KwikBucks: Correlation Clustering with Cheap-weak and Expensive-strong Signals. In N.S. Moosavi, I. Gurevych, Y. Hou, G. Kim, Y.J. Kim, T. Schuster, & A. Agrawal (Eds.), *Proceedings of The Fourth Workshop on Simple and Efficient Natural Language Processing* (pp. 1–31). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.sustainlp-1.1

Törnberg, P. (2023). ChatGPT-4 Outperforms Experts and Crowd Workers in Annotating Political Twitter Messages with Zero-Shot Learning. *arXiv*. https://doi.org/10.48550/arXiv.2304.06588

Wang, L., Yang, N., Huang, X., Yang, L., Majumder, R., & Wei, F. (2023). Improving Text Embeddings with Large Language Models. *arXiv*. https://doi.org/10.48550/arXiv.2401.00368

Watters, C., & Lemanski, M.K. (2023). Universal Skepticism of ChatGPT: A Review of Early Literature on Chat Generative Pre-trained Transformer. *Frontiers in Big Data*, *6*. https://doi.org/10.3389/fdata.2023.1224976

Wlezien, C. (2005). On the Salience of Political Issues: The Problem with "Most Important Problem." *Electoral Studies*, *24*(4), 555–579. https://doi.org/10.1016/j.electstud.2005.01.009

Zhang, T., Ladhak, F., Durmus, E., Liang, P., McKeown, K., & Hashimoto, T. (2023). Benchmarking Large Language Models for News Summarization. *Transactions of the Association for Computational Linguistics*, *12*, 39–57. https://doi.org/10.1162/tacl_a_00632

Appendix A

Table A1. Cluster Coherence Assessment Guidelines Scheme

Levels	Definitions and examples
Level 0: No Coherence	**Definition**: The two links have nothing in common. **Example**: One link discusses an environmental policy regarding renewable energy, while the other covers a new education curriculum in schools. These stories do not share thematic elements.
Level 1: Broad Thematic Coherence	**Definition**: The two links pertain to the same broad area of politics (e.g., economy, health, taxes, immigration, environment, safety, ...) but refer to stories with different actors, events, places, or organizations. **Example**: Both links cover economic issues. One is about tax reforms affecting small businesses, and the other discusses federal spending on infrastructure. They share a broad theme of economic policy but focus on distinct topics.
Level 2: Specific Thematic Coherence	**Definition**: The two links have specific actors, events, places, or organizations in common but refer to different journalistic stories. **Example**: Both stories mention the World Health Organization's response to health crises but from different angles—one focuses on funding and resource allocation, while the other examines the impact of WHO guidelines on national health policies.
Level 3: Same Journalistic Story	**Definition**: The two links refer to the same journalistic story, covering the same actors, events, places, and organizations with closely related narratives. **Example**: Both links detail discussions and outcomes of a specific international climate summit, including the same participating countries, agreed-upon actions, and criticisms from environmental groups.
Level 98: I do not know/I am not sure/one or both links contain multiple themes/stories	**Definition**: The coder is unable to assess the coherence (either because of a lack of knowledge or because of the nature of the content). **Example**: At least one of the link titles or descriptions do not clearly convey its topic to the coder. |

Table A2. Labels accuracy assessment guidelines scheme

Levels	Definitions and examples
Misfit	**Definition**: The label fails to align with the item's content, missing significant aspects or inaccurately representing its implications. **Conditions**: If the label fails to meet either the Thematic Alignment or Implications criteria, it is automatically categorized as a Misfit. **Criteria Examples** 1. Thematic Alignment: The label introduces themes or subjects completely absent in the item. 2. Implications: The label implies a stance or narrative that contradicts the item's factual content or focus.
Partial Fit	**Definition**: The label relates to the item in terms of theme and implications, but either the content covered by the label is too narrow or broad (include also other distinct themes not discussed by the item), or its context diverges from the context of the item. **Conditions**: The label meets the criteria for Thematic Alignment and Implications but fails to completely cover Content Coverage and/or Contextual Alignment. **Criteria Examples** 1. Thematic Alignment: The label addresses the key theme of the item. 2. Implications: The label accurately reflects the item implications. 3. Content Coverage: The label encompasses themes or details that extend beyond the scope of the item or cover part of the item's content well but overlooks or inaccurately represents other significant parts. 4. Contextual Alignment: The label fails to reflect the item's specific geographical, cultural, or situational context accurately.
Good Fit	**Definition**: The label fully and accurately represents the item across all aspects. **Conditions**: The label must completely satisfy all four criteria: Thematic Alignment, Content Coverage, Contextual Alignment, and Implications. **Criteria Examples** 1. Thematic Alignment: Addresses the main themes or significant content of the item clearly. 2. Implications: Accurately reflects the implications or conclusions supported by the content. 3. Content Coverage: Captures all critical details, with only minor aspects possibly overlooked. 4. Contextual Alignment: Fits well within the context presented in the item with only slight inaccuracies.

Table A3. Prompts we used to feed gpt-4-turbo

Role	Message
System	"You are an assistant tasked with aiding a political scientist in analyzing social media content related to the [2018/2022] Italian elections. Your objective is to synthesize the core themes of groups of politically themed links shared on Facebook into succinct, descriptive labels in English. These labels should encapsulate the primary themes, issues, or narratives prevalent among the links in each group, providing a concise overview of their collective content."
User	"Presented below is a selection of links from one such group. Each entry merges the title and description of a link, offering a glimpse into its thematic content. Based on these summaries, identify and articulate overarching themes or characteristics shared across these links. Your response should be a concise, descriptive phrase or label that accurately captures these shared elements. This label will be instrumental in cataloging and analyzing the political discourse related to the [year] Italian elections on Facebook. [text] (Placeholder where the actual items to be analyzed are inserted.) What descriptive English label best summarizes these shared characteristics?"

Giada Marino – Department of Communication Sciences, Humanities and International Studies, University of Urbino Carlo Bo (Italy)
https://orcid.org/0000-0002-9087-2608 | giada.marino@uniurb.it
https://aoir.social/@GiadaM
Giada Marino holds a Ph.D. from the University of Urbino Carlo Bo (Italy), where she is currently a postdoctoral researcher in the vera.ai project. Her research focuses on the intersection of information disorder and political polarization, with a particular emphasis on how citizens engage with political content and discussions on social media platforms. She uses a mixed methods approach to analyze these dynamics.

Fabio Giglietto – Department of Communication Sciences, Humanities and International Studies, University of Urbino Carlo Bo (Italy)
https://orcid.org/0000-0001-8019-1035
https://aoir.social/@fabiogiglietto
Fabio Giglietto is an Associate Professor of Internet Studies at the University of Urbino Carlo Bo, holding a Ph.D. from the same institution. He explores the complex relationship between information theory, media, and digital technologies and their impact on society. His research delves into how these forces shape social systems and influence public opinion and political communication. Giglietto also employs advanced computational methods to analyze online information dissemination and media manipulation.

https://doi.org/10.6092/issn.1971-8853/19566
Sociologica. V.18 N.2 (2024)
ISSN 1971-8853
https://sociologica.unibo.it/

AI Methodology Map. Practical and Theoretical Approach to Engage with GenAI for Digital Methods Research

Janna Joceli Omena* [a] Antonella Autuori [b, c] Eduardo Leite Vasconcelos [d]
Matteo Subet [b] Massimo Botta [b]

[a] Department of Digital Humanities, King's College London (United Kingdom)
[b] University of Applied Sciences and Arts of Southern Switzerland (SUPSI)
[c] RMIT University, Melbourne (Australia)
[d] Universidade Federal da Bahia (UFBA) (Brazil)

Submitted: May 15, 2024 – Revised version: September 5, 2024
Accepted: September 13, 2024 – Published: October 30, 2024

Abstract

This essay accounts for a novel way to explore generative artificial intelligence (GenAI) applications for digital methods research, based on the AI Methodology Map. The map is a pedagogical resource and a theoretical framework designed to structure, visually represent, and explore GenAI web-based applications. As an external object, the map functions as a valuable teaching material and interactive toolkit. As a theoretical framework, it is embodied in a static representation that provides principles for engaging with GenAI. Aligned with digital methods' practical, technical, and theoretical foundations, the map facilitates explorations and critical examinations of GenAI and is supported by visual thinking and data practice documentation. The essay then outlines the map principles, its system of methods, educational entry points, and applications. The organization is as follows: First, we review GenAI methods, discussing how to access them, and their current uses in social research and the classroom context. Second, we define the AI Methodology Map and unpack the theory it embodies by navigating through the three interconnected methods constituting it: making room for, repurposing and designing digital methods-oriented projects with GenAI. Third, we discuss how the map bridges GenAI concepts, technicity, applications and the practice of digital methods, exposing its potential and reproducibility in educational settings. Finally, we demonstrate the AI Methodology Map's application, employing a digital methodology to analyze algorithmic race stereotypes in image collections generated by nine prominent GenAI apps. In conclusion, the essay unveils methodological challenges, presenting provocations and critiques on repurposing GenAI for social research. By encompassing practice, materiality and theoretical perspective, we argued that the AI Methodology Map bridges theoretical and empirical engagement with GenAI, serving them together or separately, thus framing the essay's main contribution. We expect that the AI Methodology Map's reproducibility will likely lead to further discussions, expanding those we present here.

Keywords: Generative Artificial Intelligence; GenAI; Digital Methods; AI in Education; Image Networks; Technicity; Algorithmic Race Stereotypes.

* J.J.Omena@kcl.ac.uk

Acknowledgements

This essay is situated within the context of the "Designing With: A New Educational Module to Integrate Artificial Intelligence, Machine Learning and Data Visualization in Design Curricula" research project, that aims to develop and define a new educational module suitable for multidisciplinary environments to integrate Artificial Intelligence (AI), Machine Learning (ML), and Data Visualization (DV) into Design curricula. The research project was financially supported by the International Program of Movetia (https://www.movetia.ch), from September 2022 to February 2024. Movetia promotes exchange, mobility, and cooperation within the fields of education, training, and youth work — in Switzerland, Europe, and worldwide. We extend our gratitude to the project team for their contributions throughout the development of this work. A special thank you to all the students who participated in the workshops; their engagement and insights were invaluable. Thanks to Nerea Calvillo for contributing her insightful feminist perspective and comments to the case study on algorithmic race stereotypes. We also wish to acknowledge the 2023 Generative Methods Conference of Aalborg University in Copenhagen, where we first introduced the AI Methodology Map and its results.

We sincerely thank the two blind reviewers for their invaluable insights and comments. Additionally, we acknowledge the use of GenAI methods for proofreading this article.

1 Introduction

This essay accounts for a novel way to explore generative artificial intelligence (GenAI) based on the AI Methodology Map[1]. The map is a pedagogical[2] resource (interactive toolkit and teaching material) and theoretical framework designed to structure, visually represent, and explore generative artificial intelligence (GenAI) web-based applications (apps) for digital methods-led research. In particular, the explorations of apps and code-based platforms mediating access to GenAI foundation models (Burkhardt & Rieder, 2024). The map is an interactive toolkit and teaching material to support workshops and AI sprints, and it is also embodied in a static representation which covers theoretical orientation principles for engaging with GenAI. While we

1. The map is available at https://genmap.designingwithai.ch/map and documented at https://github.com/zumatt/AI-Methodology-Map. The AI Methodology Map integrates an experimental and multidisciplinary ongoing project, namely "Designing With: A New Educational Module to Integrate Artificial Intelligence, Machine Learning and Data Visualization in Design Curricula". It is a research project in collaboration between the Institute of Design, SUPSI; the Universidade NOVA de Lisboa, iNOVA Media Lab, and the EPFL.

2. The term pedagogical refers to the theoretical-practical framework based on Bloom's Taxonomy (Anderson & Krathwohl, 2001), reflecting the educational approaches, practices and purposes that should characterise education in the 21st century.

expect the reader to take these perspectives together, they can also serve separate purposes if desired.

The AI Methodology Map is based on three core principles: the theoretical and practical foundations of digital methods (Marres, 2017; Omena, 2021a), visual thinking and documentation of data practices (Arnheim, 1980 & 2001; Mauri et al., 2020), and interdisciplinary research efforts (Gray et al., 2022). Unlike method protocols and recipes that present "how to" steps to achieve a specific research outcome while ensuring reliable results (see Bounegru et al., 2017), the map prioritizes ways of knowing GenAI. That is understanding *what to look at* when leveraging GenAI to advance digital methods. Therefore, the map expands established digital methods practices, i.e., enacted by the repurpose of crawling, scraping, and API calling for social and cultural research, by enquiring and experimenting with *what counts in practice* when repurposing GenAI.

The AI Methodology Map differs from quick responses to the AI impact and (mis)uses with precautionary measures, as it is not focused on mandating transparent disclosure of the use and performance of large language models (LLMs) (see Stokel-Walker & Noorden, 2023; Dwivedi et al., 2023) or promoting a framework that primarily centres on the ethical issues and misuses of GenAI in educational settings (see Russel Group, 2023; Popescu & Schut, 2023; Baidoo-Anu & Ansah, 2023). Although acknowledging these as critical factors, we argue that the effort to understand GenAI from uncomplicated and technical perspectives — as the map proposes — is equally relevant. The map, thus, addresses other challenges of "repurposing" GenAI (technology) for social research, which involves most of all, a mindset (see Franklin, 1990; Marres, 2017) encompassing conceptual, technical, and empirical dimensions (see Hoel, 2012; Omena, 2022; Rieder, 2020). By creating space for GenAI to sit through hands-on practice, the map aims to surface foundational layers in discussions for social research and contributes to the field of digital methods epistemology.

This essay outlines the AI Methodology Map principles, its system of methods, educational entry-points, and applications. The organization is as follows: First, we review GenAI methods, discussing how to access them and their current uses in social research and the classroom context. Second, we define the map and unpack the theory it embodies, navigating through the three interconnected methods constituting it: *making room for Generative AI* (method 1); *repurposing GenAI apps and outputs* (method 2); and *designing digital methods-oriented projects with GenAI outputs* (method 3). Method 1 focuses on ways to become familiar with GenAI conceptually, technically and empirically. Method 2 introduces new ways to use GenAI and repurposing prompting techniques as research methods. Method 3 elicits the exploration of designing digital methods projects for analyzing GenAI models, outputs, or interfaces. Third, we discuss how the map bridges GenAI, technicity, applications and the practice of digital methods, demonstrating its potential and reproducibility in three educational settings. Finally, a case study demonstrates the AI Methodology Map's application, employing a network vision methodology (Omena, 2021b) to analyze image collections generated by nine prominent GenAI apps. This study investigates algorithmic race stereotypes and compares visual models' responses to the same prompt. We conclude by discussing methodological challenges and addressing three provocations.

This essay's main contribution is the introduction and development of the "AI Methodology Map", a dual-purpose interactive toolkit and theoretical framework designed for exploring GenAI applications in digital methods-led research within the Social Sciences and Humanities. By functioning as both a theoretical framework and a practical tool, the map bridges a gap between theoretical perspectives and empirical engagement with GenAI, and facilitates its integration into educational and research contexts.

2 Generative AI Methods: From Definition and Accessibility to Social Research Applications and Classroom Context

GenAI is a subset of machine learning (ML) that employs deep generative models to generate novel and realistic content across various modalities (e.g., text, images, code) based on user prompts[3]. To facilitate user interaction with such models, interfaces are developed as tools that use prompts as interaction touchpoints (Banh & Strobel, 2023). Each model necessitates different types of input data and is enabled to generate specific outputs, exemplified by the functionality of input data to output data, which may include text-to-text, text-to-image and other operations. The development of generative AI is contingent upon the integration of three essential components: a dataset utilized in the training of the large language model (LLM), the source code employed to define and execute the training process on a given dataset, and the model eventually comprising the parameters or weights (Shrestha et al., 2023).

The ability of GenAI models to produce *previously unseen synthetic content* (García-Peñalvo & Vázquez-Ingelmo, 2023) differs from classification tasks performed by predictive ML models, such as identifying constitutive elements and semantic contexts in an image, e.g. person, woman, happy. GenAI models offer unpredictable synthetic content. On the one hand, the meaning of language is created through the user inputs (data or prompt) and the model's capacity to recognize existing information and generate new content (Gozalo-Brizuela & Garrido-Merchan, 2023). On the other hand, the specificity of GenAI models can shape research methodologies as what they generate exposes their internal knowledge space (see Borra, 2024; Burkhardt & Rieder, 2024).

2.1 Accessing GenAI Methods through Web Apps and Coding Platforms

The accessibility of LLMs to generate content — identified in the essay as GenAI methods — may be achieved through two distinct modes, as shown in Figure 1. One can access GenAI models through open source or proprietary (1) web applications and (2) coding platforms, which allow us to carry out tasks using the model in different ways yet requiring different skill sets.

GenAI web applications offer intuitive interfaces requiring no prior technical knowledge, such as Dall-E 2 (Ramesh et al., 2022) or ChatGPT (OpenAI, 2023) for generating images and text. That is, one accesses the GenAI methods via front-end interface interactions only. Other examples are research software web-based applications, such as Prompt Compass (Borra & Plique, 2024)[4] which provides access to various LLMs, offering a library of prompts for digital research and allowing users to apply these prompts to a series of inputs. GenAI coding platforms allow interaction with the model through code, providing more customization and often more control over the data used. This generally requires medium to high programming skills. Models can be accessed via global information trackers like GitHub or coding platforms like HuggingFace. Examples include Meta Llama 2 (Touvron et al., 2023), multiple large language models (LLMs) that have already been trained and refined, and Stable Diffusion (Rombach et al., 2021), a model that can be used to generate or modify images based on text prompts.

3. That is a piece of text or input provided to a GenAI model which directs and shapes the model's response.
4. https://github.com/ErikBorra/PromptCompass

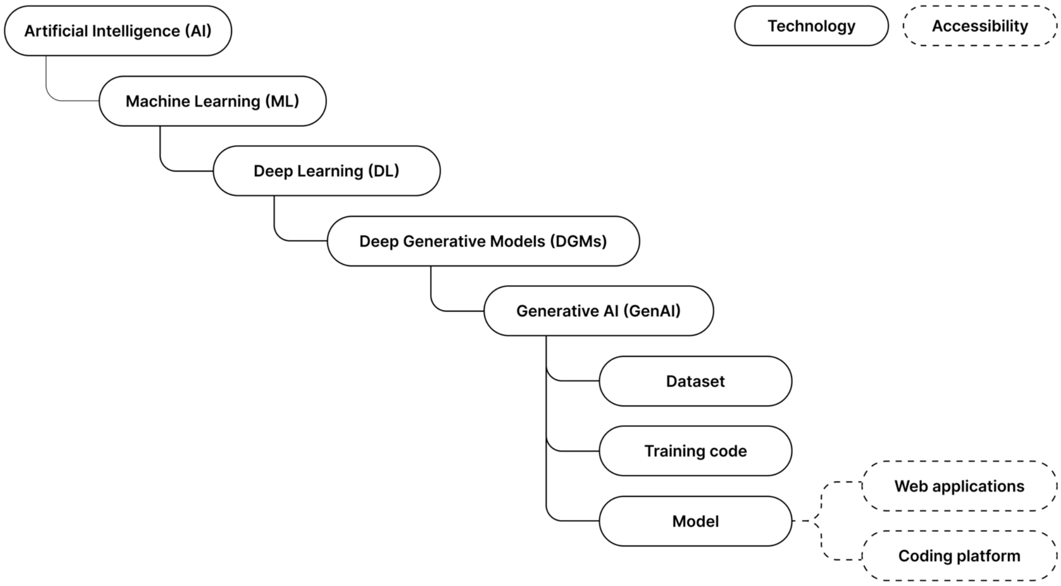

Figure 1. The placement of Generative AI in the realm of Artificial Intelligence (Inspired by Banh & Strobel, 2023) including GenAI components and the model accessibility

Discussions about the disparities between GenAI web applications and code-based platforms involve their accessibility and management of model settings. GenAI web applications are typically proprietary software that does not permit open access, limiting control over model settings. In contrast, code-based applications are often released as open-source software. This allows users unlimited access to their models, trained datasets, and code for personalized training (Shrestha et al., 2023). This provides users greater control over the usage of GenAI models. However, the coding nature of the software could be a limitation. To overcome this gap, some open-source solutions use libraries, such as Gradio (Abid et al., 2019), that allow developers to create quick demos or web applications. This openness fosters a collaborative environment where developers and researchers contribute to the model's improvement, leading to more robust and refined AI applications. Such models are typically made available to the public with comprehensive documentation, facilitating customization and experimentation for specific research or project requirements. However, compared to web applications, they still require expertise in understanding and managing the code source.

The proposed AI Methodology Map uses an interactive visualization[5] that groups several different GenAI apps and coding platforms, both proprietary and open-source, allowing the discovery and exploration of generative methods.

2.2 GenAI and Social Research

The integration of AI in the field of social sciences has led to significant changes in the approaches and methodologies used in research (Sinclair et al., 2022), providing insights into human behaviours, social patterns (Zajko, 2021), online communities, hidden dynamics, data interpretation (Koplin, 2023) and enhancing collaboration between humans and machines across

5. http://genmap.designingwithai.ch

https://doi.org/10.6092/issn.1971-8853/19566

various fields (Perez et al., 2023). Greene (2023) and Anderson et al. (2023) demonstrate how the combinations of AI technologies with traditional research methodologies are shaping new areas of knowledge and understanding of the social domain. Such convergence is indeed stimulating the development of inventive ideas and protocols, transforming the already recognized and employed methodologies and providing new perspectives on human behaviours, on the social patterns inscribed within AI systems, which can reveal societal trends and intricate phenomena (Leshkevich & Motozhanets, 2022).

The employment of generative AI in social research not only opens up new ways of understanding and addressing social challenges (see Wang et al., 2022) but unveils new ethical concerns (Graziani et al., 2023) and implications for decision-making processes, educational frameworks, and interdisciplinary cooperation. In research practices, generative AI leads to merging data, algorithms, and social practices, giving life to new and unexpected cultural phenomena and dynamics (de Seta et al., 2023). There is a call for ongoing dialogue and collaboration among scholars, policymakers, and education officials to discuss the incorporation of (generative) AI into research and social applications (Graziani et al., 2023). Potential alternatives include developing AI ethical guidelines sensitive to cultural nuances (Vogel, 2021) and fostering interdisciplinary collaboration to enrich the debate and develop new methods to effectively address and mitigate AI bias (Ferrara, 2024).

Building upon the characteristics and potentials of GenAI, which is efficient in analyzing and discovering social patterns and behaviours, as well as perpetuating inequalities, misrepresentations, or distortions of physical reality inherent to the training datasets, interdisciplinary research such as media studies, design, and digital methods have leveraged GenAI. This has often involved encoding stereotyped representations of society to expose existing biases (Luccioni, 2023). An emerging practice of repurposing GenAI outputs has proven valuable and may be considered for social research. Generating images for further scrutiny using qualitative and quantitative methods (Venturini, 2024) is one example. Results have shown that Stable Diffusion often amplifies racial and gender disparities, especially concerning job occupations (Nicoletti & Bass, 2023). The analysis of over 5,000 images reveals that white males tend to occupy leadership roles while associating people of colour with lower-paying jobs or criminal activity. When it comes to the depiction of biodiversity across GenAI models, it varies by language, model, and context, with notable consistencies and differences (Colombo, De Gaetano & Niederer, 2023). Language and model choice significantly affect the representation of species and human presence. Seasonal and geographical prompts influence the colour scheme and thematic focus, while ecosystem and continent prompts highlight the challenges in accurately depicting biodiversity, sometimes making it stereotypical, decorative, and simplified. Instead of analyzing collections of generated images, Erik Salvaggio (2023) employs media studies approaches to qualitative interpretations as reflections of cultural, social, economic, and political biases. A generated image of a kissing couple — showing a white heterosexual couple with the man appearing reluctant and distorted — reveals underlying assumptions about gender, intimacy, and representation. He suggests that understanding the dataset's origin, content, and collection method is crucial for uncovering the biases encoded in AI-generated images, providing insight into societal norms and values. These cases underline how scholars from an interdisciplinary background are using GenAI outputs as valuable perspectives to expose the need for inclusive and culturally sensitive GenAI development practices.

2.3 GenAI in Classroom Context

GenAI models are challenging educational institutions with entirely different modes of operation and knowledge production from what we have seen so far. After causing a combination of shock and hysteria (Goulart, 2024), GenAI has already transformed traditional teaching methodologies due to its capacity to *impact*, *modify*, *and enhance* students' performance and learning experiences — particularly since 2022 and after the public can easily access GenAI web applications, like Midjourney's open beta version in July and OpenAI's ChatGPT in November. Discussions within higher education institutions and scholarly literature have explored the integration of GenAI web applications into pedagogical practices (Honig et al., 2023; Russel Group, 2023); such as educational curricula, pedagogical strategies, and assessment methodologies must be reevaluated or are already being redesigned and created (see Botta et al., 2024; Verhoven & Vishal, 2023; Antolak-Saper et al. 2023). Higher education institutions response to GenAI, such as those in Australia, Brazil, Spain, Portugal, and the United Kingdom, have been majorly inclusive and welcoming, yet a more practical approach is still under development (see Antolak-Saper et al., 2023; Agência Lusa, 2023; Gaspar, 2023; Roussel Group, 2023). Examples include adopting experiential teaching methods, developing critical field guides, designing GenAI in teaching planning and classroom activities, and creating new methodological frameworks.

Verhooven and Vishal (2023) advocate for "experiential teaching methods", emphasizing skills such as emotional intelligence, collaboration, creativity, and critical thinking — attributes that machines cannot easily replicate. Their perspective underscores the necessity to equip students with competencies that are indispensable in the dynamic and technology-driven job market of the future. Honig et al. (2023) discuss three ways of applying GenAI to teaching methods. Firstly, AI can assist students during the ideation phase and help them explore solutions and problems. Secondly, it can be a peer reviewer in code development and improve software maintainability. Thirdly, and aligning with the Socratic method, AI can actively participate in discussions, fostering critical thinking through inquiry and debate. This last role covers two essential learning outcomes: identifying misinformation and developing skills in using AI tools consciously.

Critical field guides also explain new educational formats to account for AI in a classroom context, such as the "Critical Field Guide for Working with Machine Learning Datasets" (Ciston, 2023). This guide promotes critical thinking by introducing the conscious use of AI, providing straightforward technical definitions — e.g. models, neural networks — and emphasizing the importance of understanding the ecosystem behind the Graphical User Interface (GUI) of AI applications. This includes identifying the creator/s of the dataset, the labelling method, the types of data contained, the contexts included, the state of updating and documentation of the dataset, the licenses and terms of use, the people or groups of people involved and interested in the dataset, and so forth. Only through this process of continuous inquiry into the model that a technical and ethical awareness can be developed to access these datasets as valuable resources for designing outcomes.

Finally, projects proposing new frameworks and didactic guidelines with AI integrate innovative learning experiences for students. This is the case of "Designing With: An Educational Module to Integrate Artificial Intelligence, Machine Learning, and Data Visualization in Design Curricula" (Botta et al., 2024). The project proposes a design-stage-oriented framework and didactic guidelines tailored explicitly for design students and teachers. The framework integrates design stages with AI and data visualization tools, enabling students to explore collaborative opportunities in a structured and informed way.

GenAI in the classroom context cultivates critical and creative thinking and analytical skills among students when interacting with AI-generated outputs. As predicted by Gordon Pask (1975), a British cybernetician and inventor, interactions with machines indeed enable us to exchange and learn while reflexively reshaping our knowledge bases through iterative questioning and critical engagement. This essay contributes to this moment by offering hands-on educational methods that touch upon foundational GenAI aspects.

3 The Map: An Introduction

This section introduces the AI Methodology Map, represented in Figure 2, as a theoretical framework that outlines principles of orientation for engaging with GenAI. It offers the map's definition, unpacks the theory it bears, and navigates three interconnected methods to understand, explore, and develop digital methods projects with GenAI.

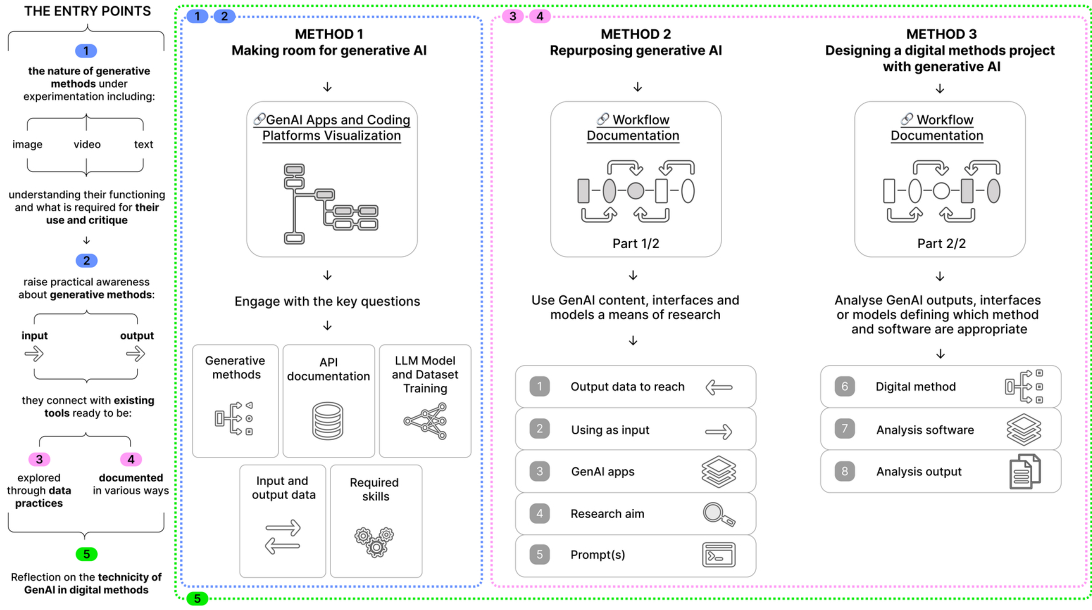

Figure 2. The AI Methodology Map: A static representation that provides principles and methods for engaging with GenAI. The map's digital and interactive version is available at https://genmap.designingwithai.ch/map

3.1 AI Methodology Map: What Is and What For?

The AI Methodology Map (Figure 2) is a pedagogical resource (interactive toolkit and teaching material) and theoretical framework designed to structure, visually represent, and explore GenAI web-based applications for digital methods-led research. The map is a conceptual, empirical and interactive structure that organizes knowledge and methodological frameworks for engaging with GenAI. It combines methods crafted to enhance comprehension of GenAI through practical applications that help researchers and students develop ways of understanding, thinking about, and creating knowledge using GenAI. Theoretically, it covers perspectives and discussions on empirical engagement with GenAI in the Social Sciences and Humanities.

As an external object (digital version), the map is materialized as teaching material and an interactive toolkit for exploring GenAI Apps in the context of digital methods research. While we expect the reader to take these perspectives together, they can also serve separate purposes if desired.

The methodology map presumes that users possess a basic understanding of generative methods. Its representation serves not only as a visual guidance for practical activities but also aims to make the "invisible" (see Mauri & Ciuccarelli, 2016) aspects of GenAI methods more visible and understandable. Thus, this map has a different purpose from method protocols and recipes, which record and present predefined, structured methods, techniques, or procedures designed to achieve a specific research outcome (Bounegru et al., 2017; Mauri et al., 2020). While method protocols ensure the reliability and validity of empirical findings (Cross, 2001), explaining method design and implementation (what and how it was done), the map we introduce prioritizes processes of acquiring technical knowledge for method reasoning and practicing. By focusing on "what to look at", the map elicits ways of *knowing GenAI while understanding when and why to value them in a methodological ensemble* (see Omena, 2021a). In this sense, the map's purpose and outputs move towards the epistemology of digital methods and its critical reflections rather than final research products.

Regarding reproducibility, although the map is developed for implementation in workshops or AI sprints, it allows anyone to independently repeat the procedures without needing mediators. Individuals can use the map's essential theoretical points as a guide to engage with GenAI and take advantage of the external teaching resources. In the following sections, we will introduce the theoretical framework that underpins the map and its system of methods, which elicit attitudes of making room for, repurposing, and designing projects with GenAI.

3.2 Three Principles: Theoretical Framework

Three principles underlie the AI Methodology Map: (i) the practical and theoretical foundations of digital methods, (ii) visual thinking, and data practice documentation, along with (iii) interdisciplinary research endeavours. We will discuss each individually and then illustrate how they intersect within the interconnected methods depicted on the map.

The map embodies a technicity perspective on the practice of digital methods that considers medium-technicity to (re)think the design and implementation of these methods (Omena, 2021a; 2022). This perspective attends to developing a specific mindset, modes of thinking, and technological awareness required by the methods (Marres, 2017; Rogers & Lewthwaite, 2019; Rogers, 2013) — or technology itself (see Franklin, 1990), yet it embodies a domain of knowledge encompassing conceptual, technical, and empirical dimensions (see Hoel, 2012; Rieder, 2020) about GenAI and the necessary computational media requested to work with the methods. On the one hand, a technicity perspective is closely related to relational processes between the researcher and the computational media required to advance the methods, i.e. the iterative and navigational research practices that constitute a methodological ensemble, technical and practical knowledge. On the other, it refers to the researcher's attitude to understanding GenAI and computational media conceptually, technically, and empirically, in isolation and comparison and on their terms, while knowing how and when to appreciate their substance, value and agency (Omena, 2021a; 2022). This framework, as elucidated in the AI Methodology Map, encourages a grasp of GenAI methods and applications on their terms and relational contexts within a methodological ensemble, as demonstrated in section 5.

The second principle underlying the map incorporates visual thinking and data practice documentation to acquire and produce knowledge about GenAI. Visual thinking on the map guides intuitive and intellectual modes of thinking that closely interact, making it difficult to separate them (see Arnheim, 1980; 2001). The *map's visual representations* support processes of visually acquiring knowledge through thought and experience. They are designed to connect the map user with core aspects of GenAI and its exploratory applications.

Visual thinking in the map is not just a feature but a comprehensive approach to *introducing* and *revealing* GenAI through the context of digital methods practices. For example, it allows users to easily navigate three interlinked methods that offer detailed procedures and a clear set of instructions for overcoming the challenges of repurposing GenAI for social research. This approach, where the process of acquiring and producing knowledge involves the interpenetration of theoretical, practical, and technical modes of thinking, is further enhanced by integrating *visual data practice documentation*. The latter aids in recognizing the "non-objective, situated, and interpretative nature" of data practices (Mauri & Ciuccarelli, 2016; Mauri et al., 2020). For example, the map guides users in structuring and recording each step and decision via methodological workflows. The visual aids are particularly relevant as they facilitate a practical and technical understanding of GenAI apps, encouraging critical, reflective, and relational thinking. Visual thinking is also applied through *interactive visualization*, which provides an initial technical and practical knowledge of GenAI apps, models, or code.

The third principle explains how the map fosters interdisciplinary research efforts that combine digital methods, information design, and media studies. The AI Methodology Map is based on research-led teaching (see Gray et al., 2022; Rogers & Lewthwaite, 2019) and collaborative approaches among SUPSI, NOVA, and EPFL that involve MA courses such as Interaction Design, New Media and Web Practices, Space and Communication, and an MSc in Transition, Innovation, and Sustainability Environments. The proposed methodology combines the authors' research background and classroom context to advance research while teaching about GenAI, its potential for design and media studies, and its use as a research method.

Together, these principles correspond to and inform five practical entry points for leading the map's application (see Figure 2), which promote understanding and engagement with GenAI apps. We will discuss and illustrate them practically in Section 4.

3.3 Three Methods: Make Room, Repurpose, and Design Projects with GenAI

The AI Methodology Map combines three interconnected methods designed to understand, explore, and develop projects with GenAI (Figure 2). These methods follow a technicity perspective to engage with GenAI (Omena, 2021a) and are better suited for individual and group activities in AI sprints or workshops.

Making room for GenAI (Figure 2, Method 1) explores generative methods by navigating an interactive visualization[6] while responding to crucial questions about GenAI apps, supporting the map user's conceptual, technical, and empirical familiarization with them. The interactive visualization contains structured information about various generative methods mediated by GenAI proprietary applications and open-source models. Whereas five key questions ask what generative method and what LLM is operating. Also, is API documentation available, and can we identify the dataset used to train the model?[7] What are the limitations or potential

6. https://genmap.designingwithai.ch/

7. As Borra (2024) explains, "foundation models are (pre-)trained on massive data sets — and are mainly probabilistic completion machines. Fine-tuned models use foundation models as their basis, but have learned to do specific tasks such as classification, extraction and summarisation."

biases one might encounter in the LLM currently in use? Is it an open-source or proprietary model? Who developed it? What type of input is required? What kind of output does one get? What is required to use this app or open-source code? If possible, adjust the model temperature; what does it mean? The explorations and findings should be documented in a shared file[8] (e.g. using Figma), which allows for and empowers collective discussions among all involved. The proposed activities encourage efforts to become acquainted with GenAI methods as carriers of meaning — here, employing GenAI for social research. Method 1 showcases that to know GenAI apps or open-source models, one must do more than interact with them. So, when making room for GenAI, the initial fascination with its methods is immediately balanced with a critical and technical awareness of what they are and the key elements making them operate.

Repurposing GenAI (Figure 2, Method 2) *for social research or media research* is a method that creates new ways of using GenAI and prompting engineering techniques without fundamentally changing their nature (see Rogers, 2013; Noortje, 2017). In other words, the creative use of prompting and their outputs, GenAI apps' interfaces or code as research methods or objects of critique. *Repurposing* refers to established digital methods practices for conducting research using materials not initially created or intended for that purpose, such as digital objects (hashtags, URLs, web entities) and web technologies and methods (crawlers, scrapers, APIs, knowledge graphs). Sections 4 and 5 demonstrate how GenAI models and generated images can be repurposed to uncover racial stereotypes. Repurposing GenAI is an extension of method 1: because I now understand GenAI, I will take a risk in repurposing it.

The map user engages with a rationale that intentionally starts with medium specificity and only then defines the research aim accordingly. Once the generative method(s) and associated web-based application or open-source model are defined, we determine the expected outputs and required inputs. For example, text, instructions, or tables could be used to generate audio, but which of these options is most compatible with the attitude of *repurposing* GenAI for social research? What are the reasons behind that choice, or why not opt for a given input? Then, one tries, tests, and generates prompts while "being mindful of prompt formulation" as their different settings can shape the outcomes (see Borra, 2024). Examples involve creating research personas, using search queries (see Colombo et al., 2023; Borra, 2024) and political positioning efforts (see Hartman et al., 2023; Rozado, 2023) as prompts or involving specified and underspecified prompts to capture gender bias in the LLMs training datasets, as we demonstrate in section 5. The decisions made are visually documented in a shared file, allowing all parties to see how generative methods are being repurposed.

Designing digital methods projects with and about GenAI (Figure 2, Method 3) organizes a workflow responsive to Method 2 and open to experimental and exploratory analysis of GenAI models, outputs, and interfaces. It is a way to explore new forms of knowledge production. As an extension of the previous methods, now: because I understand what aspects of GenAI can be repurposed, I will design a digital method project with it. Once again, decisions are recorded in a shared file. Many questions arise about what we should look at and how to implement methods, such as how to analyze GenAI visual, textual, and audio outputs. This essay does not answer these questions directly but illustrates possibilities mapped by applying the AI Methodology Map in research-led teaching and learn-by-doing workshops (see section 4). It also showcases that GenAI visual-generated content can be repurposed with digital methods research (see section 5).

8. https://genmap.designingwithai.ch/teaching-resources

4 The Map's Applications: From Technical Awareness to Social Investigations

This section introduces the AI Methodology Map as an interactive toolkit and teaching material. It describes three situations in which the map's theoretical perspective is applied in practice, and how empirical engagement informs its theory. Using a research-led teaching approach, we integrated existing studies in digital methods, communication design, and media studies to shape the workshop content. This approach facilitated the exploration of GenAI apps and encouraged students to critically engage with these topics. Master's students participated in hands-on workshops and AI sprints[9], where they learned by actively working with the GenAI apps, AI concepts and research software.

The map's first application focused on applying conceptual principles, what we called getting familiar with GenAI, a five-day workshop to integrate GenAI methods into design practices. The second application, a six-hour workshop, focused on exploring and repurposing GenAI apps for social research. We created an environment that allowed master students to expand their methodological imagination to address social, political, cultural, or environmental issues while critically examining AI models. In the third application, we developed a study to investigate algorithmic race stereotypes in the context of image generation using digital methods.

We argue that the map's application (Figure 2) bridges GenAI concepts, technicity, and the practice of digital methods by intentionally crafting interconnected methods that raise conceptual awareness about GenAI while technically and empirically engaging with it. Differing attitudes focus on how we respond quickly to GenAI's impact with preventive measures; the map takes a step back, slowing down reactive practices while creating spaces and opportunities for GenAI to sit. First, reflecting an awareness component about the generative method and the AI platform mediating access to the LLMs. That is a vision of GenAI from both conceptual and technical perspectives. Second, investing in the specific mindset to work with digital methods and GenAI while accounting for relational processes inherent to these methods is something that only unfolds in technical practice. That is a vision of technicity due to the inevitable proximity or a particular relationship we must develop with computational media and AI necessary to implement the method (Omena, 2021a). The Map then advances a technicity perspective, operationalized through educational entry points (Figure 2, colour highlights) and cultivates an awareness component about GenAI and its potential for and as research methods

- The nature of the generative method under experimentation
- The essential inputs and outputs of these methods
- Data practices
- Data documentation
- The technicity of GenAI in developing digital methods-oriented projects (and vice-versa)

The educational entry points in the Map play a crucial role in promoting understanding and engagement with GenAI. They are designed to showcase the potential of methods that are co-designed with and about generative AI. The entry points for data practices and documentation

9. The student sample was defined according to the author's institutional affiliations and teaching agenda.

are intentionally created to encourage users to ask relevant questions about generative methods, understand the role of prompts, and acknowledge the mediating role of other analysis software and the researcher's intervention in interpreting GenAI outputs, models or interfaces.

Finally, the map's theoretical framework and methods were employed in three educational contexts, demonstrating its potential and reproducibility. These applications helped refine the map's methods and visual documentation and are described and discussed in the following sections to support the argument that the AI Methodology Map can bridge GenAI, technicity, applications, and the practice of digital methods.

4.1 First Application: Getting Familiar with (Generative) AI

The first application occurred in July 2023 during a one-week workshop[10] at the University of Applied Sciences and Arts of Southern Switzerland (SUPSI) entitled "Designing With: AI, ML, DV", involving 18 multidisciplinary students and workshop facilitators from different fields of design, architecture and social sciences[11]. The educational experience was organized into two modules, covering other AI methods rather than exclusively GenAI methods. The first module, *Getting Familiar With*, aimed to provide students with the basic theoretical and practical skills of AI, ML, and data visualization (DV) through the introduction to literacy and guided practical exercises with applications specific to each discipline. The second one, *Get in Depth With*, aimed to support students, divided into multidisciplinary groups, in developing and practicing the methodology for designing with AI.[12]

During the workshop, students were provided with three design challenges to start exploring and exploiting the framework and choosing the most appropriate AI web applications or open-source models to employ. The design challenges addressed different topics and fields of research, such as "Designing for Digital Twin Cities", "Designing for Digital Interactions" and "Designing for Social Phenomena" (Figure 3). Group work was supported by workshop facilitators according to their interests and research competencies in the field. Additionally, throughout the workshop, students were asked to document the integration of AI, ML, and DV tools within each stage of the design process, keeping track of steps and choices. The process of documenting was intended first to foster the acquisition of a method, second, it allowed a holistic analysis during the evaluative phase of the workshop, enabling a thorough examination of the methods and frequencies at which students systematically exploited and integrated these tools (Figure 3).

10. This workshop was developed in the context of research project "Designing With: A New Educational Module to Integrate Artificial Intelligence, Machine Learning and Data Visualization in Design Curricula" (Botta et al., 2024). It supported testing and validation of the "Designing With Interactive Framework", accessible at the link https://designingwithai.ch/interactive-framework.

11. The "Designing With: AI, ML, DV" workshop included six students of the SUPSI Master of Arts in Interaction Design, four students of NOVA Master in New Media and Web Practice, two students of the NOVA Master of Science in Transition, Innovation, and Sustainability Environments, two students of the HEAD Master in Space and Communication, and three students of the EPFL Master in Architecture. The workshop was part of a broader research, founded by Movetia in 2021, entitled Designing With A New Educational Module to Integrate Artificial Intelligence, Machine Learning and Data Visualization in Design Curricula, in collaboration between the SUPSI Institute of Design, the Universidade NOVA de Lisboa and the EPFL (École polytechnique fédérale de Lausanne) Media x Design Lab. The website of the full project is accessible via https://designingwithai.ch/.

12. The AI Methodology Map, conceptualized before the workshop and as the inspiration for its modules, has since been further expanded with a specific focus on generative AI web applications for digital methods-led research.

BRIEF	Designing for Digital Interactions		Designing for Social Phenomena		Designing for Digital Twin Cities	
PROJECTS	Dew	Shift	Political Activism	Tomato Girl Summer	Monolith to faceless	Smart Move
Understand	ChatGPT, Elicit, Visualcrossing	/	/	/	Google Maps	ChatGPT, Google Maps
Define	NotionAI, ChatGPT	/	Down Them All!	Down Them All!, Phamtombuster, Zeeschuimer, 4CAT, Google Sheet	Vision API	/
Ideate	ChatGPT	/	/	/	Midjourney	ChatGPT
Prototype	TwoTone, ML5.js, PoseNet, RunwayML	Melobytes, Teachable Machine, ML5.js, P5.js, RunwayML, Poly AI	Vision API, Memespector, Gephi, Figma, Label studio, Magic AI, Typeset.io	Vision API, Memespector, Gephi, Rawgraphs, Imagesorter, Voyant Tools, Midjourney, Teachable Machine	Imagery, Midjourney	ML5.js, P5.js, COCO SSD, Midjourney
Develop	ChatGPT, Github	Melobytes, RunwayML, Garagebands	Table2net	Teachable Machine, Gephi, Midjourney	Imagery	Midjourney, Blace, After Effects
Release	Github	/	/	/	/	/

AI Tools ML Tools CV/DV Tools Other Tools

Figure 3. Visual comparison of the Artificial Intelligence, Machine Learning, and Computer Vision tools employed by student groups in their projects at various stages of the design process during the Getting in Depth module of the Designing With AI, ML, and DV workshop

We selected two projects to provide a comprehensive overview of the framework usage, according to different briefs, AI applications employed, project objectives, and the nature of the final artefacts. The first project, "Dew" (Amietta et al., 2023), developed by Raffaele Amietta, André Filipe Nunes Matos, and Adèle Guilbault, explores innovative data representation and interaction via generative AI models, enabling new communication forms between machines and users, transforming perceptions of machine-human interactions in a combined digital-physical realm. It originated from an inquiry posed to ChatGPT (OpenAI, 2023). The query addressed was:

> Student: "*What is the difference between a human and a computer?*"
>
> ChatGPT: "*Humans exist in the physical world, with sensory perception. Computers are digital entities that do not have a physical presence. They interact with the world through input devices, but they lack sensory experiences*"

Starting from this answer, students formulated a set of research questions: "How can computers have a sensorial perception of the world? And how can we see or hear what a machine is perceiving from the data it's collecting? How can we interact with it together?".

These inquiries led to the development of a novel digital interface, "Digital Embodiment Wave" (DEW), that shifts the perspective from human to machine, using collected data from the weather (temperature, humidity, wind speed, solar radiation) to generate new music. DEW communicates what it observes from the data being fed into it, generates sound, and allows for interaction with humans through gestures, enabling the creation of new collaborative outputs.

Students started collecting weather datasets from the 18th of July from Visual Crossing (Visual Crossing Corporation, 2003). The data collection process yielded a spreadsheet containing columns representing different aspects of weather, including temperature, real feel, dew point, humidity, wind gusts, wind speed, cloud cover, and solar radiation. The data was subsequently translated into sound compositions utilizing TwoTone (Rogers, & Cairo, 2022), a GenAI app dataset-to-sound. To make data experienceable, a user-machine interaction prototype was developed utilizing ML5.js and PoseNet (Kendall et al., 2016) to facilitate real-time gesture-based sound modulation. Additionally, Runway ML (Valenzuela et al., 2018) was employed to generate the videos serving as the visual assets of the user-experience interface. Lastly, to enhance the immersive quality of the experience, a web application was created and hosted at https://nerd-life-squad.github.io/about, in collaboration with GitHub and the code-assistant capability of ChatGPT (Figure 4).

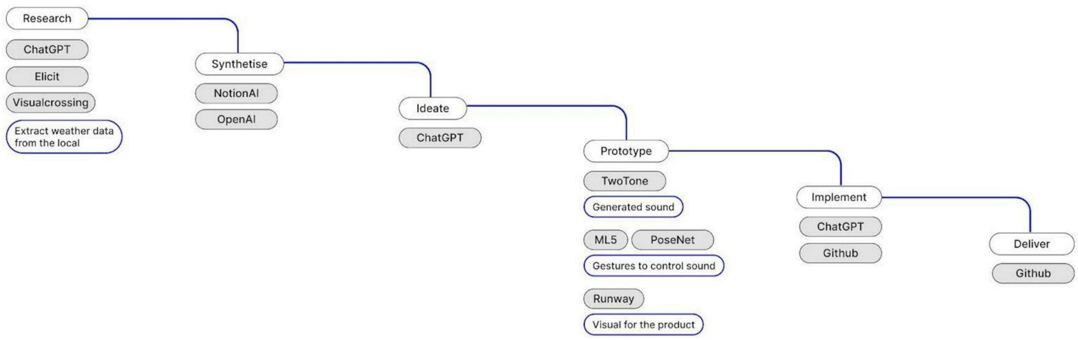

Figure 4. "Dew" project (Amietta et al., 2024). The protocol diagram illustrates the AI/ML/DV tools employed at each stage of the design process

The second project, "Tomato Girl Summer", developed by Catherine Yu, Jean Louise Tschanz-Egger, and Mariana Souto (2023), explores the imagery of the social media trend tomato girl across various platforms. The project integrates and mixes several AI/ML/DV applications at different stages of the design process, by performing digital methods-led research (Figure 5). The research phase involved gathering data — including images, captions, comments and likes — from social media platforms TikTok and Instagram by using various tools like DownThemAll (Maier et al., 2004), PhantomBuster (Boiret, 2016), Zeeschuimer (Peeters, 2023), 4CAT (Peeters & Hagen, 2022), and Google Sheets. Then, collected data was processed using Vision AI to detect web entities. For this purpose, Memespector GUI (Chao, 2021) and Google Vision were applied to recognize the visual content and text associated with the images. Subsequently, Gephi (Bastian et al., 2009), Rawgraphs (Mauri et al., 2017), Image Sorter (Visual Computing Group, 2018), and Voyant Tools (Sinclair & Rockwell, 2003) were employed to create various types of data visualizations, such as networks of image descriptions, word clusters, engagement graphs, and colour grids, representing the different findings of the analyses. The final stages involved prototyping and implementing interactive outputs based on the research insights. This included, first, generating ideal images of tomato girls, perhaps as archetypes or examples of the trend, using the Midjourney (Midjourney Inc., 2022) text-to-image model. Then, a machine learning model was trained with the Teachable Machine application (Google Creative Lab, 2017) to recognize the characteristics of a "tomato girl" using the images generated. Lastly, this model was used to analyze live camera feeds from users interacting with the interface, enabling real-time identification and interaction with the "tomato girl" aesthetic and visual phenomena.

This project demonstrates how the integration of AI and ML applications with digital methods-led research aims to render the results of analyses more tangible and experiential. This approach enhances user engagement by fostering empathy with the social phenomena under study. For example, the model trained with Teachable Machine enables users to embody themselves with the tomato girl visual trend and connect more personally with the research concept. Furthermore, the utilization of Midjourney, employed to generate images of women that adhere to conventional gender norms, has the potential to prompt new inquiries into how text-to-image models comprehend and contribute to the perpetuation of such stereotypical imagery.

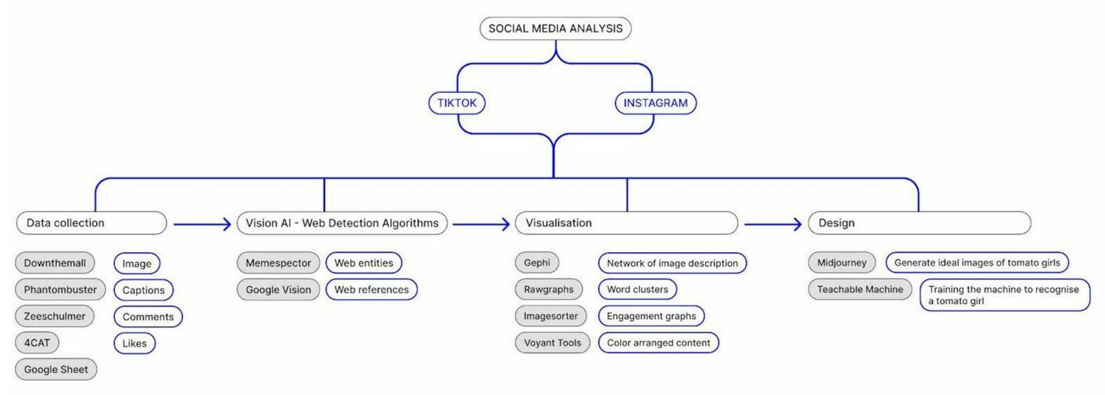

Figure 5. "Tomato Girl Summer" project (Yu et al., 2023). This diagram illustrates the protocol of the process, combining digital methods research with design methodologies to explore and analyze the "tomato girl" imagery across social media platforms

At the end of the workshop, the students completed an evaluation questionnaire concerning the content and pedagogical activities. The feedback gathered, together with the project outcomes designed by students, proves the framework's comprehensibility, effectiveness, and potential.

4.2 Second Application: Generative AI as Research Methods

The second application occurred in October 2023; a six-hour workshop titled "Generative AI as Research Methods" took place at Universidade Nova de Lisboa as part of the Erasmus Mundus MA in Transition, Innovation, and Sustainability Environments. Before the workshop, students participated in seminar sessions to learn about generative AI and image-generation methods. The first part of the workshop involved individual exploration of various generative methods using the GenAI interactive visualization, followed by group discussions to generate collective ideas on leveraging GenAI as a research method. The discussions revolved around the possibility of repurposing generative methods outputs to study or identify socio-technical, cultural, or political issues. Participants also discussed which GenAI method(s), such as Midjourney for image and Copilot for text generation[13], and why use them. The second part focused on group work, with students working on their chosen generative method and receiving project guidance. The project documentation was designed on Figma. The final output was a detailed description of the workflow (Figure 6), elucidating decisions on implementing digital methods on the generative method outputs.

The workflows mainly focused on exploring prompting techniques across models for text generation (e.g. ChatGPT), image generation for understanding how models are fed (e.g. algo-

13. Two broad options were suggested. Social research for mapping social, political, cultural, or environmental issues, or medium research-oriented project to interrogating generative methods via their outputs.

rithmic bias detection), and advancing app walkthrough methods for audio and video generations. For example, a group of students designed a comparative analysis of audio generation apps (Murf vs ElevenLabs[14]) with a specific focus on accent, tonality, the overall quality of the audio, and gender variations and using text input to generate audio about the weather warning alert in Portugal (see Figure 6). They observed that Murf (Edkie et al., 2020) offers consistent audio outputs in terms of tonality and accent for a single voice actor, while ElevenLabs (Dąbkowski, & Staniszewski, 2022) had inconsistent output using speech synthesis.

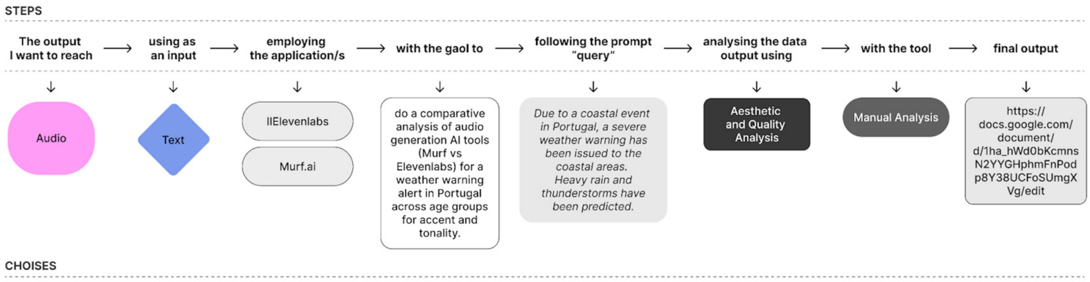

Figure 6. Methodological workflow documentation of the Audio Generation project. Group members: Jan Carlo M. Castro, Ayesha Zulfiqar, Shivam Shumsher, David Vuth

Some methodological limitations in quickly accessing audio, image, and video generation models due to GenAI apps' pricing plans, or difficulties in automatically analyzing textual outputs from prompts using ChatGPT (OpenAI, 2023), have led to the conclusion that traditional digital methods may not always be the most effective solution. However, exemplary projects resulting from the six-hour workshop, present innovative ways of repurposing GenAI outputs for social research. One example is the "Situating Generative-AI Pain & Pleasure" project, developed by MSc's students Jan Carlo M. Castro and Shivam Shumsher. The project interrogates how Craiyon (Dayma, 2022), a free image generator app, represents pain and pleasure. Sixty-four prompts (32 for pleasure, 32 for pain) were divided into four categories: senses, population, ethnicity, and continents. Sub-categories were created for each category. Using the Digikam software (The DigiKam Team, 2001), they tagged 640 images with labels describing entities, emotions, ages, and genders for both pain and pleasure.

Overall, the project findings (Figure 7) reported that Craiyon depicts pain predominantly through human forms with a consistent red spot. Pain images often feature blue translucent human-like abstract subjects and are more associated with older adults and children, particularly among migrant and refugee populations. Pleasure representations show growing diversity, mainly featuring human subjects and a broader colour spectrum. Pleasure images are more likely to depict adults, particularly women, with children often shown in brighter backgrounds. Emotionally, pain images are frequently labelled with sadness and are significantly associated with refugees, migrants, and older adults. Pleasure images are more varied, with happiness, disgust, and sadness almost equally represented, often showing children and older adults with smiling faces. Gender-wise, pain images show a mix of genders with a higher representation of females and abstracted non-binary figures. Pleasure images are predominantly female, with even higher percentages when specific ethnic groups or professions like sex workers.

14. https://murf.ai/; https://elevenlabs.io/

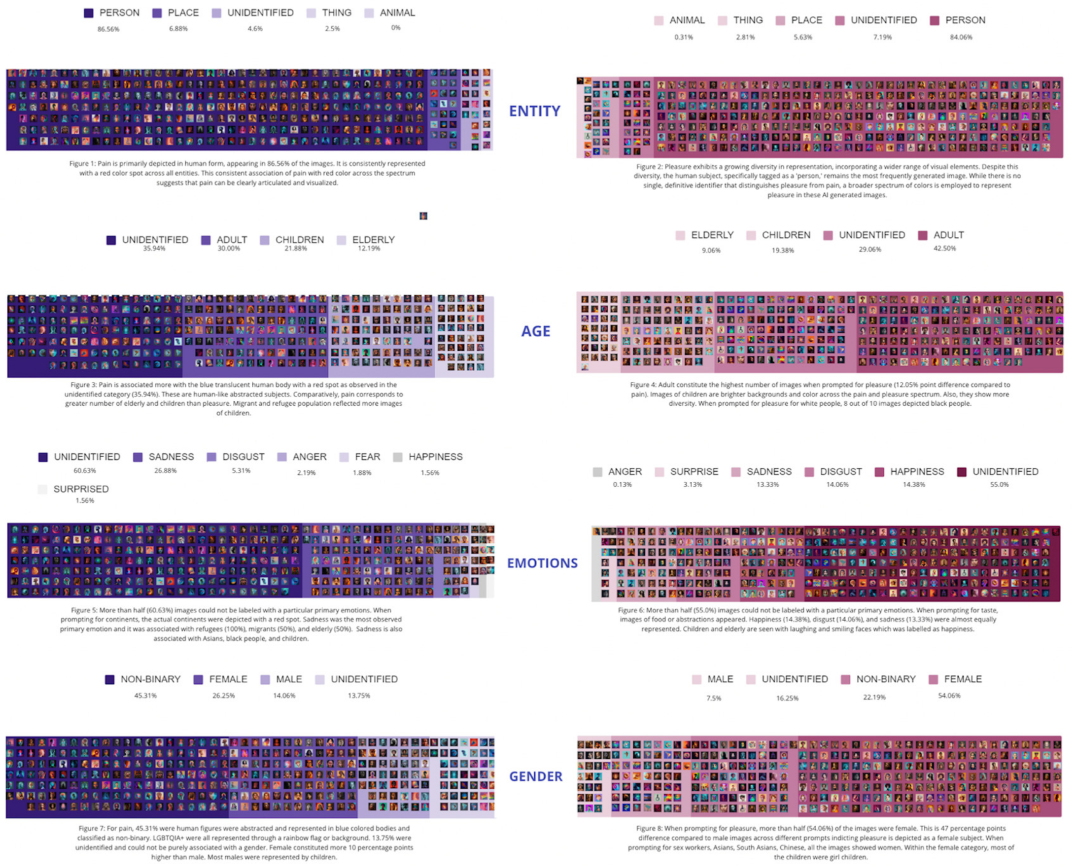

Figure 7. Exemplary project resulting from the Generative AI as Research Methods workshop at Universidade NOVA de Lisboa. On the right, image collections represent *pleasure*; on the left, *pain*. Situating Generative-AI Pain & Pleasure project developed by Jan Carlo M. Castro, and Shivam Shumsher. Image source: Castro & Shusher, 2023. Miroboard link: https://miro.com/app/board/uXjVNCx_UhM=/

4.3 Third Application: Repurposing GenAI Visual Outputs

We repurpose GenAI to critically reflect on methods to study the inherent bias constituting image generation models (see Sun et al., 2023; Chauhan et al., 2024; Gorska & Jemielniak, 2023). Two prompting techniques were adopted to create image collections with nine popular image generation models (Figure 8). We used both underspecified and specified prompts, the latter being a more detailed instruction that specifies elements such as the style of the image (e.g., pop art), the subject's positioning within the frame, background details, and more. Secondly, we transferred conventional digital methods for interpreting GenAI-generated content, such as building, visualizing, and narrating computer vision networks (Omena et al., 2021) and arranging images by model, prompt, and hue with an image montage (Manovich, 2020).

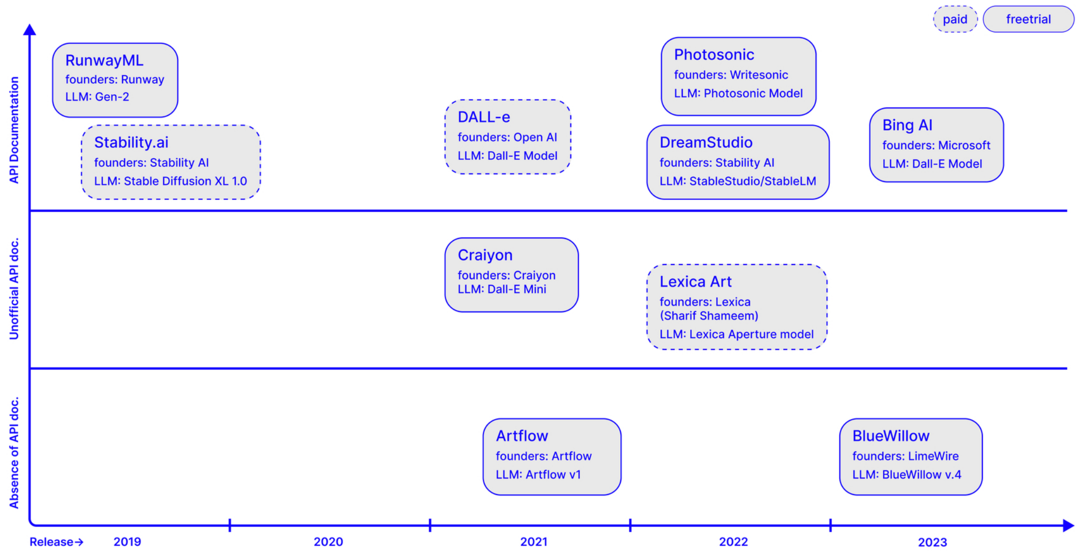

Figure 8. Overview of the GenAI apps used in the Case study

The study's main objective was to critically interrogate GenAI visual outputs and detect algorithmic race stereotypes in the context of image generation. Results showed that out of nine models, only Microsoft Bing AI would produce images associating Black women with violence and guns. Also, the nine GenAI apps (Artflow, Bing AI, Craiyon, DreamStudio, RunawayML, Dall-E, Lexica, and Stability.ai) tend to depict Black Women-as-sign trope (see Báez, 2023): Black women are either portrayed as young, skinny and beautiful or associated with serious and angry expressions. The case study was motivated by the public report of a Brazilian State Deputy, Renata Souza, who got a racist output from Microsoft Bing AI in October 2023, as we will detail and discuss in the next section.

5 The Workflow: How Can GenAI Visual-generated Content Be Repurposed Using Digital Methods?

This section describes how we repurposed GenAI methods and visual outputs from nine models to expose racial stereotypes. It begins by situating a case study triggered by the generated content of a Black woman holding a gun, despite the original prompt specifying a Disney Pixar-like image. This raised the question of to what extent the current dominant GenAI apps for image generation (see Figure 8) contribute to the perpetuation of Black women's detrimental stereotypes. We then explain the digital methods and network vision methodology[15] used and conclude by discussing the main preliminary findings.

5.1 A Black Woman Movie Star in the Favela – GenAI's Biassed Take with a Gun!

On October 25th, 2023, Renata Souza, a state deputy from Rio de Janeiro, shared a video on Instagram (Figure 9) that exposed a racist issue with the Microsoft Bing AI generative models. Souza created a prompt to generate a Disney Pixar movie poster with her as the leading role. Her prompt was based on the following instructions:

> A Disney Pixar-inspired movie poster with title "Renata Souza". The main character is a Black woman using afro hair tied up, dress an African style blazer. The scene should be in the distinct digital art style of Pixar, a favela in the back, with a focus on character expressions, vibrant colors, and detailed textures that are characteristic of their animations, with the title "Renata Souza" (Souza, 2023).

The model generated an image of a Black woman holding a gun with a favela in the background (Figure 9). She expressed her surprise and outrage in an Instagram video[16], stating that she had never mentioned weapons or violence in her instructions. She had only requested a poster featuring a Black woman in a favela, but the model added a gun to the image. "This is proof that algorithmic racism exist!", she said. This GenAI output exposes algorithmic bias and discrimination embedded in Microsoft Bing AI models and how the data they were trained on reveals past discrimination, i.e. the association of a Black woman in a favela with violence, a poor performance when generating images of underrepresented groups (Buolamwini, 2017). It also reflects central discussions on algorithmic discrimination and oppression ingrained in artificial intelligence technologies (Noble, 2018; Sharma, 2024) despite these issues being obscured by the rhetoric of technology's neutrality (Noble, 2013). The context of the case study uncovers how Bing AI is perpetuating these patterns. After the repercussion, Bing AI blocked the prompt used by Souza, arguing that it "might be in conflict with our content policy". Despite that, the model had no problem generating images when we excluded Souza's name from the prompt.

As argued by Kassom and Marino (2022), social researchers may not only account for technical understandings but also consider "the broader social impact of an algorithm's use and whether that use contributes to or ameliorates racial inequity" (p. 2). Reports of AI bias, discrimination, and misleading or poor specific cultural representations from proprietary AI have been well documented by researchers from diverse backgrounds (see Birhane, 2022; Buolamwini & Gebru, 2018; Silva, 2023). Examples include Google Photos tagging Black people as gorillas in 2015, Stable Diffusion associating Black men with gang members in 2022, Midjourney failing to generate images of Black doctors treating white children in 2023, and Canvas feature marking Black hairstyles as insecure in 2024 (Silva, 2023). By repurposing

15. Regarding reproducibility, the network vision methodology was developed by Janna Joceli Omena and her collaborators (see Omena et al., 2021). This methodology is currently under formalization. The step-by-step process can be easily repeated by anyone, including those not familiar with digital methods, by following this document: https://docs.google.com/document/d/e/2PACX-1vR8IZJKni6j1tG8KE872LS8HsqBVe-PKSIlqVG5mMAfR7vUKTzmW_T9TPSe7mA-GVwroLwMS5I96dbq/pub. Further discussion on these methods is available at Omena, 2021b.

16. https://www.instagram.com/reel/Cy1p6EQpwXB/?igshid=MzRlODBiNWFlZA

Figure 9. Brazilian deputy Renata Souza's Instagram post contains both the prompt she used on Bing AI and the output the application generated. Source: Souza, 2023

GenAI apps and associated LLMs for image generation, this case study joins efforts in documenting GenAI race stereotypes in the context of image generation, having as a starting point Renata Souza's Disney Pixar movie poster by Bing AI's biassed take with a gun (Figure 9).

5.2 Designing Digital Methods Research with GenAI Visual Outputs

To what extent do current dominant GenAI models for image generation contribute to the perpetuation of Black women's detrimental stereotypes? To answer this question, we employed network vision methods (Omena et al., 2021) to visualize and analyze images generated by nine GenAI apps and associated LLMs with computer vision and through networks (Figure 10). In other words, we repurpose GenAI visual outputs to (1) investigate the response variations among generative models when presented with identical prompts and (2) examine the presence (or absence) and characteristics of racial stereotypes, particularly associations between Black individuals and violence, across different models. Thus, the main objectives of the case study are not only to investigate GenAI-related social issues but also to interrogate generative models' outputs, considering that these outputs "are not simply lookups or search queries over the training data" but an entry to access the "transformer intelligence" (Burkhardt & Rieder, 2024, p. 4) of the AI image generation models.

Considering the unpredictability of GenAI outputs and the models' constant updates based on users' practices (Burkhardt & Reider, 2024), we first conducted tests prompting various image-generation models to explore and compare results. Next, we defined the formulation of specified and underspecified prompts (see Figure 10). The former reproduces the original prompt by the Brazilian deputy so that we could assess how nine GenAI apps respond to it. The latter adapts the original prompt into a broader one, reducing it to its main keywords for a deeper scrutiny of the models' responses: "A movie poster starring a Black woman".

Making image collections is the second step. We generated 30 images for each prompt using Artflow (Wojcicki, 2020), Bing AI (Microsoft, 2023), BlueWillow (Limewire, 2023), Craiyon (Dayma, 2022), Dall-E 2 (Ramesh et al., 2022), Dream Studio (Rombach et al., 2021), Lexica (Shameemm, 2022), RunwayML (Valenzuela et al., 2018), and Stability.ai (Mostaque, 2019). We paid US$ 40 to generate images with Dall-E 2, Lexica, and Stability.ai. Methods for visual-

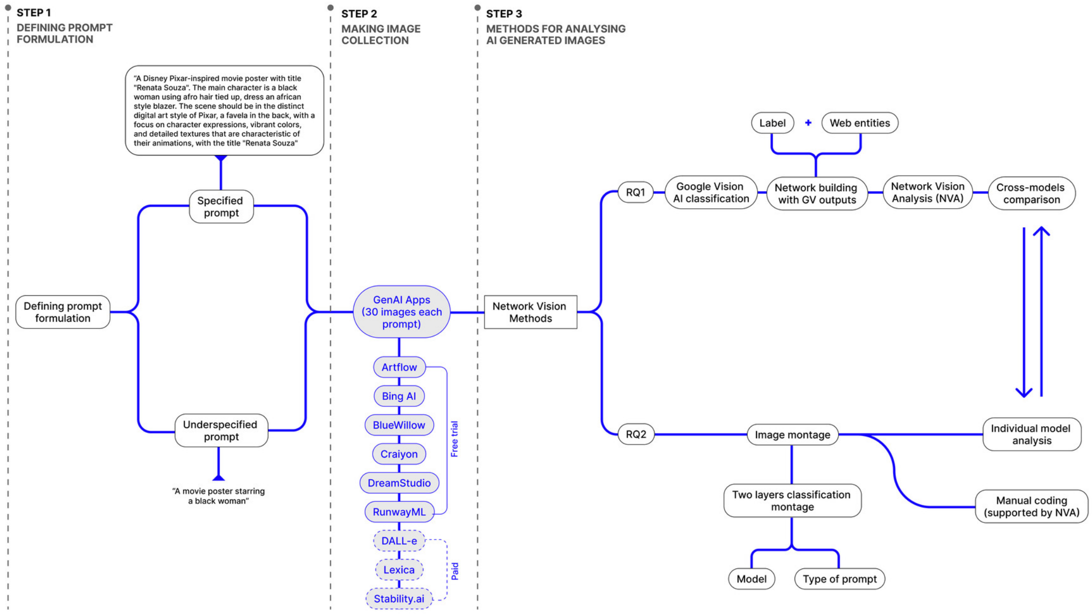

Figure 10. The method protocol for repurposing GenAI is to detect and expose racial visual stereotypes

izing and analyzing the generated 540 images are the third step. We built networks to examine patterns, similarities, and specific characteristics in portraying Black women across GenAI apps and their responses to the same prompts. Additionally, we arranged images according to each GenAI app's output, prompt, and image hue. Both methods, separated and complementary to each other, helped us identify racial stereotypes responsive to our specified and underspecified prompts.

The methods employed in this study can serve as a template for investigating and addressing various social issues by repurposing GenAI outputs and examining generative models' responses. They can be adapted for other projects exploring different societal concerns.

5.3 Findings: Visual Biases in AI-Generated Content

Overall, Bing AI is the only model displaying images of Black women associated with violence and guns, with four instances featuring guns in its underspecified prompt images. A lack of body diversity was detected, with a recurring pattern depicting Black women as young and slender. Another prevalent stereotype pertains to facial expressions, where a serious, angry, or intense gaze is commonly attributed to Black women. When it comes to generative visual models, such as those used by Craiyon, results show they were not trained to capture specific cultural and contextual nuances, such as accurately representing a Brazilian "favela". In Brazil, favelas are urban environments inhabited by low-income communities and are often associated with social challenges such as poverty, violence, and lack of adequate infrastructure. This result exposes the need to include cultural sensitivity and proper training to ensure that GenAI models accurately capture and represent various social and cultural contexts. Below, we present detailed findings based on the research questions.

RQ1: "How do different visual generative methods respond to the same prompt?"

The network vision analysis revealed that most models respond similarly to both prompts. Images generated by Bing AI, BlueWillow, DreamStudio, Lexica, RunwayML, and Stability.ai were positioned by ForceAtlas2 (Jacomy et al., 2014) in the centre of both networks. Therefore, these models' images were tagged mainly with the same labels/web entities by Google Vision AI, i.e., respond to prompts with the same image styles. With the specified prompt (Figure 11), these models generated 3D cartoon images in a medium close-up shot portraying a Black woman facing forward with afro hair in vibrant attire against colourful backgrounds. For the underspecified prompt (Figure 12), these models mainly generated medium close-up shot images of Black women facing forward with afro hair, this time with a photography style and orange/brown colour palettes.

Three models stood out of the majority when responding to the prompts, with their images placed in the periphery of both networks. Craiyon only generated solid colour backgrounds in both prompts; therefore, its imagery was mainly tagged with labels/web entities related to facial features and hairstyles. That is significant for the specified prompt because it explicitly asked for a favela background, to which Craiyon did not respond. Artflow's images, however, tend to have more complex backgrounds than the other models' outputs. Because of that, in both prompts, its images look more like a film frame than a movie poster. Finally, Dall-E 2 generated 2D cartoon images in both prompts, showing an imperfect reproduction of the Pixar-like style demanded in the specified prompt and distancing its imagery from other models that mainly generated photographic images when responding to the underspecified one.

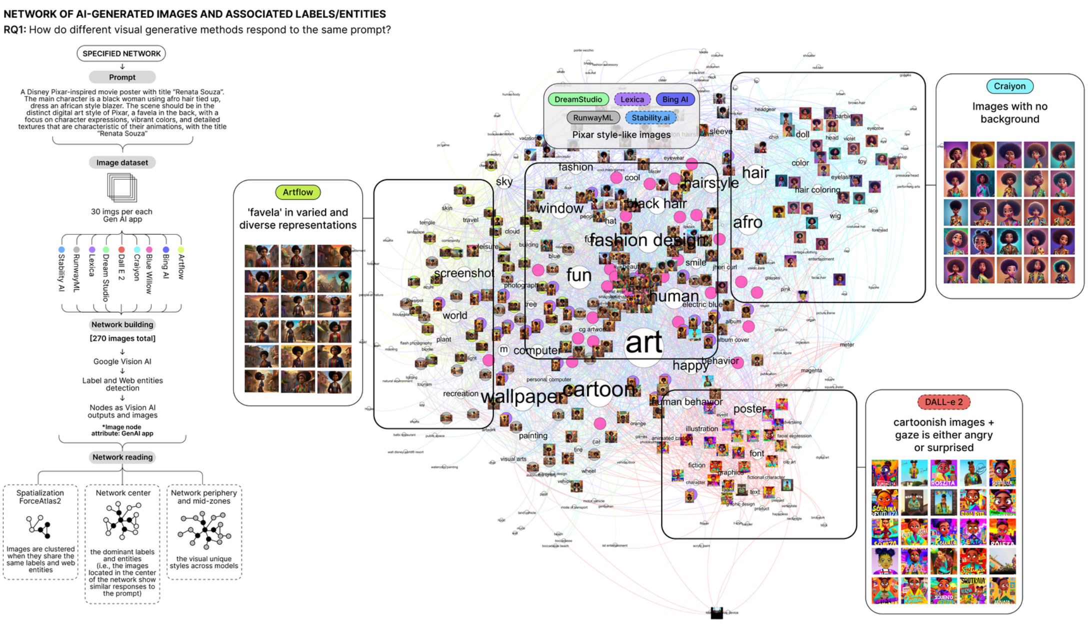

Figure 11. Network of AI-generated images and associated labels and web entities. This computer vision network derives from the **specified prompt** and associated images interpreted by Google Vision's label and entity detection. Node positions indicate image clusters based on the co-occurrence of computer vision labels and entities classifying one or more images.Node colours represent the GenAI apps that generated these images.

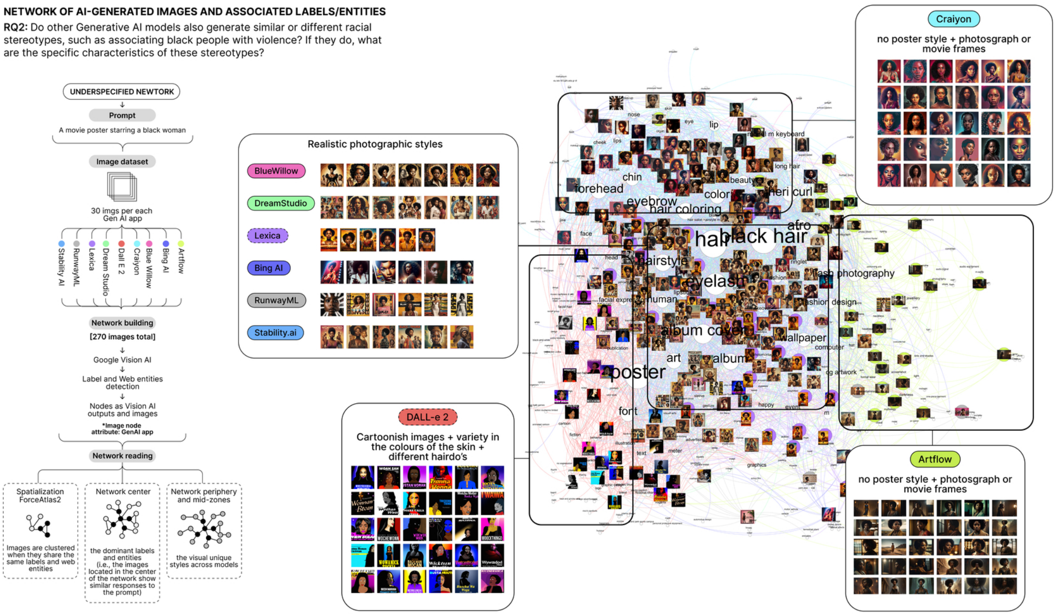

Figure 12. Network of AI-generated images and associated labels and web entities. This computer vision network derives from the **underspecified prompt** and associated images interpreted by Google Vision's label and entity detection. Node positions indicate image clusters based on the co-occurrence of computer vision labels and entities classifying one or more images. Node colours represent the GenAI apps that generated these images.

RQ2: *"Do other Generative AI models also generate similar or different racial stereotypes, such as associating Black people with violence? If they do, what are the specific characteristics of these stereotypes?"*

The image grid visualization (Figure 13) facilitated the identification of racial stereotypes. In line with feminist media studies, which highlight the women-as-sign trope — where women's bodies are used as icons symbolically representing specific communities (Báez, 2023) — GenAI image models often rely on stereotypes when depicting Black women. This limited representation supports and perpetuates oppression against this community (hooks, 1992). For example, Bing AI generated images depicting black women with guns; there were four of them. Additionally, all models exhibited varying degrees of other stereotypes.

Craiyon is the only model that doesn't represent the favela as a busy, dirty, poor place because it only generates images with solid colours in the background. All models lacked diversity in body types, with no images of older women and only one plus-size woman, a result aligned with the fetishization of black women's bodies identified by other researchers (see Noble 2018; 2013). Children and teenagers only appear in the images generated by the specified prompts. There was also a lack of diversity in hairstyles, with Bing AI showing the most variation, albeit primarily in images from the underspecified prompt. Finally, images depicting black women smiling were mainly associated with the specified prompt. At the same time, those without the Pixar-inspired input often featured stern or angry expressions, with Dall-E displaying this stereotype prominently.

Figure 13. The image grid visualization assists in identifying racial stereotypes across the GenAI applications.

6 Conclusion, Challenges, Provocations

In this essay, we introduced the AI Methodology Map and its theoretical principles, system of methods, and applications in educational and research settings. The map theoretically covers perspectives and discussions on empirical engagement with GenAI in the Social Sciences and Humanities. As an external object, it is materialized as teaching material and a pedagogical tool for exploring GenAI Apps in the context of digital methods research. While we expect the reader to take these perspectives together, they can also serve separate purposes if desired. The map's conceptual framework combines the practical and theoretical foundations of digital methods, visual thinking, and the documentation of data practices, along with interdisciplinary research. As a pedagogical resource and theoretical framework, the map integrates three interconnected methods and a technicity perspective that elicit attitudes of *making room for*, *repurposing*, and *designing* projects with and about GenAI. The expected results are geared toward developing a specific mindset required to advance digital methods research rather than "how-to" steps to achieve a specific research outcome, differentiating the methodology map from a method protocol or recipe. Moreover, the map's teaching guidelines highlight a fundamental aspect present in digital methods; practical, technical, and theoretical modes of reasoning interrelate with each other, not just occasionally but essentially (Omena, 2021a; 2022). Its application, thus, is an invitation to understand GenAI from uncomplicated and technical perspectives while thinking about how we can use its outputs as research material or objects of critique. As opposed to educational concerns in social research and higher education institutions, which focus on developing methodologies to neutralize bias or disclose ethical issues and misuses of GenAI — quick responses to the AI impact. In this sense, the map contributes to building literacy in GenAI by diminishing the gap of what Mercedez Bunz (2022) has addressed as a moment of a profound human misunderstanding of AI cultures.

We argued that implementing the AI Methodology Map can open up applied scenarios that account for and repurpose GenAI for social research. The map, thus, functions as a theoretical framework and a pedagogical resource (interactive toolkit and teaching material), bridging theoretical and empirical engagement with GenAI. However, there are limitations and challenges to consider. First, the three applications of the map — not a final product as it can be expanded — serve as *just* a starting point for both understanding in practice the potential of GenAI as a research method, an object of experimentation and reflecting on the epistemology of digital methods. Second, without the skills to work with GenAI open-source code platforms and global information tracker repositories, the ease of access to generative methods relies on the AI market. *It is never a problem if you pay for it*. Consequently, the absence of free trial credits to GenAI models directly influenced the student's decisions to work with specific generative methods (e.g. text and image over audio and video generation). Limiting, then, method creativity and practice. Lastly, during the workshops, more attention could have been paid to the role of foundation LLM models and the dominant models in the AI market. Likewise, despite our efforts to explain prompts and their importance, workshop participants have not paid much attention to their role in implementing the AI Methodology Map. Effective prompting techniques were refined and implemented outside of the workshop contexts. This was possible when students were given a project assignment and extra time (weeks) to develop a digital methods project with GenAI.

While this essay has illuminated the potential applications of the AI Methodology Map, such as its theoretical points as a principle of orientation into engaging with GenAI, its use in creating image collections to scrutinize AI generative models and uncover inherent bias in their training datasets, we conclude by addressing three provocations. Regarding accessing AI methods for research purposes, there are some aspects that the history of web API creation, maintenance, discontinuation, and closure have taught us. From freely and almost unlimited access to limitedly requested access according to project themes and institutions' (or scholars') prestige to finally having no other option than paying for it. Social media and Vision AI APIs are exemplary cases[17]. If advancing digital methods comes with a cost, will we be willing to pay to access GenAI models? Should we ask questions about which models are worthwhile and why? Or aren't we just replacing the old consumption impulse to access large amounts of social media data by *generating* content with GenAI and *running* models for comparison studies? The second provocation refers to repurposing GenAI with digital methods. It is already acknowledged in the AI community that generative AI methods "are essentially projecting a single worldview, instead of representing diverse cultures or visual identities"[18] (Luccioni, in an interview for Nicoletti & Bass, 2023). If all AI models have inherent biases, should we continue to identify gaps, lacks, or absences in GenAI by developing methods based on testing and experimenting with prompt modifications? Or, should we take a step back, slow down, and make room for properly learning about prompt techniques and the models themselves?

Lastly, and for the future of digital methods, to what extent are we moving towards developing more methods for dealing with GenAI data outputs, opening up a new agenda for prompting methods? For example, are we developing methods to access a foundation model's internal "knowledge space" (see Burkhardt & Rieder, 2024), where user data is no longer centred or has become secondary? We also have learned that conventional digital methods can be transferred to analyzing GenAI outputs. For instance, telling us what we already expect, AI

17. Additionally, the automation market service has, for instance, made researchers pay for these services to track, capture, and study social and political bots or analyse the impact of social media ads.

18. See also the work of Nicoletti & Bass (2023), Sun et al. (2023) and Popescu & Shut (2023) in how GenAI takes stereotypes and gender and cognitive biases.

models are trained differently; therefore they carry unique forms of discrimination, e.g. Microsoft Bing AI associates black women with violence, generating them holding a gun. How will we employ a conceptual, technical, and empirical understanding of GenAI to think about new ways to design and implement methods?

We anticipate that the AI Methodology Map's reproducibility will spur further discussions, extending the conversation we have initiated here.

References

Abid, A., Abdalla, A., Abid, A., Khan, D., Alfozan, A., & Zou, J. (2019). Gradio: Hassle-Free Sharing and Testing of ML Models in the Wild. 2019 ICML Workshop on Human in the Loop Learning. *arXiv*, 1906.02569. https://doi.org/10.48550/arXiv.1906.02569

Agência Lusa. (2023). Universidades de Portugal, Brasil e Espanha juntam-se para discutir impacto e "transição digital" como resposta aos novos "desafios. *Observador*, 8 November. https://observador.pt/2023/11/08/universidades-de-portugal-brasil-e-espanha-juntam-se-para-discutir-impacto-e-transicao-digital-como-resposta-aos-novos-desafios/

Amietta, R., Matos, A.F.N., & Guilbault, A. (2023). *DEW*. https://nerd-life-squad.github.io/about

Anderson, L.W., & Krathwohl, D.R. (2001). *A Taxonomy for Learning, Teaching and Assessing: A Revision of Bloom's Taxonomy of Educational Objectives: Complete Edition*. New York, NY: Longman.

Anderson, C., Heinisch, J.S., Deldari, S., Salim, F. D., Ohly, S., David, K., & Pejovic, V. (2023). Toward Social Role-based Interruptibility Management. *IEEE Pervasive Computing*, *22*(1), 59–68. https://doi.org/10.1109/mprv.2022.3229905

Antolak-Saper, N., Beilby, K., Boniface, B., Bui, D., Burgess, P., Cheema, A., Crocco, M., Fordyce, R., Galbraith, K., Lansdell, G., Lim, C., Moore, J., Nathania, A., Nawaz, S., Raveendran, L., Saha, T., Sapsed, C., Shannon, B., Soh, K., Swiecki, Z., Vu, T., Wagstaff, P., Wallingford, E., Wong, P., & Zaid, F. (2023). Guides for Assessment Re(design) and Reform. *AI in Education Learning Circle*. https://www.ai-learning-circle-mon.com/

Arnheim, R. (1980). A Plea for Visual Thinking. *Critical Inquiry*, *6*(3), 489–497. https://doi.org/10.1086/448061

Arnheim, R., & Grundmann, U. (2001). The Intelligence of Vision: An Interview with Rudolf Arnheim. *Cabint Magazine*, 26 April. https://www.cabinetmagazine.org/issues/2/grundmann_arnheim.php

Báez, J.M. (2023). Performing Representational Labor: Blackness, Indigeneity, and Legibility in Global Latinx Media Cultures. *Feminist Media Studies*, *23*(5), 2455–2470. https://doi.org/10.1080/14680777.2022.2056755

Baidoo-Anu, D., & Ansah, L.O. (2023). Education in the Era of Generative Artificial Intelligence (AI): Understanding the Potential Benefits of ChatGPT in Promoting Teaching and Learning. *Journal of AI*, *7*(1), 52–62. https://doi.org/10.61969/jai.1337500

Banh, L., & Strobel, G. (2023). Generative artificial intelligence. *Electronic Markets*, *33*(1). https://doi.org/10.1007/s12525-023-00680-1

Bastian, M., Heymann, S., & Jacomy, M. (2009). Gephi: An Open Source Software for Exploring and Manipulating Networks. *Proceedings of the International AAAI Conference on Web and Social Media*, *3*(1), 361–362. https://doi.org/10.1609/icwsm.v3i1.13937

Birhane, A. (2022). The Unseen Black Faces of AI Algorithms. *Nature*, *610*(7932), 451–452. https://doi.org/10.1038/d41586-022-03050-7

Boiret, G. (2016). PhantomBuster. [Software]. https://phantombuster.com/

Borra, E. (2023). ErikBorra/PromptCompass (v0.4). Zenodo. https://doi.org/10.5281/zenodo.10252681

Borra, E. (2024). *The Medium Is the Methods: Using Large Language Models (LLMs) in Digital Research*. [Keynote]. Digital Methods Winter School, University of Amsterdam, Amsterdam, The Netherlands.

Botta, M., Autuori, A., Subet, M., Terenghi, G., Omena, J.J., Leite, E., Kim, F.C. (2024). Designing With: A New Educational Module to Integrate Artificial Intelligence, Machine Learning and Data Visualization in Design Curricula. https://designingwithai.ch/

Bounegru, L., Gray, J., Venturini, T., & Mauri, M. (Eds.). (2018). A Field Guide to 'Fake News' and Other Information Disorders. Public Data Lab. https://doi.org/10.2139/ssrn.3097666

Bunz, M. [GoetheUK]. (2022). The Culture of Artificial Intelligence. Goethe Annual Lectures at the Goethe-Institut London. [Video]. YouTube, 24 November. https://www.youtube.com/watch?v=bTR6EP34W_w

Buolamwini, J.A. (2017). *Gender Shades: Intersectional Phenotypic and Demographic Evaluation of Face Datasets and Gender Classifiers*. [Doctoral dissertation, University of Missouri]. Massachusetts Institute of Technology, Cambridge.

Buolamwini, J., & Gebru, T. (2018). Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. *Proceedings of Machine Learning Research*, *81*, 77–91.

Burkhardt, S., & Rieder, B. (2024). Foundation Models are Platform Models: Prompting and the Political Economy of AI. *Big Data & Society*, *11*(2). https://doi.org/10.1177/20539517241247839

Castro, J.C.M., & Shumsher, S. (2023). Situating Gen-AI Pain & Pleasure: Interpretative Querying Approach Combining Situational Analysis with Digital Methods [Presentation slides]. Faculdade de Ciências Sociais e Humanas, Universidade NOVA de Lisboa. http://dx.doi.org/10.13140/RG.2.2.16436.67201

Chao, J. (2021). Memespector GUI: Graphical User Interface Client for Computer Vision APIs (Version 0.2.5 beta). [Software]. https://github.com/jason-chao/memespector-gui

Chauhan, A., Anand, T., Jauhari, T., Shah, A., Singh, R., Rajaram, A., & Vanga, R. (2024). Identifying Race and Gender Bias in Stable Diffusion AI Image Generation. *2024 IEEE 3rd International Conference on AI in Cybersecurity (ICAIC)*, 1–6. https://doi.org/10.1109/ICAIC60265.2024.10433840

Ciston, S. (2023). A Critical Field Guide for Working with Machine Learning Datasets. https://knowingmachines.org/critical-field-guide

Colombo, G., De Gaetano, C., & Niederer, S. (2023). Prompting For Biodiversity: Visual Research With Generative AI. Digital Methods Summer School 2023. https://wiki.digitalmethods.net/Dmi/PromptingForBiodiversity

Cross, N. (2001). Designerly Ways of Knowing: Design Discipline versus Design Science. *Design Issues, 17*(3), 49–55. https://doi.org/10.1162/074793601750357196

Dąbkowski, P., & Staniszewski, M. (2022). ElevenLabs. https://elevenlabs.io/

Dayma, B. (2022). Crayon (v3). https://www.craiyon.com/

de Seta, G., Pohjonen, M., & Knuutila, A. (2023). Synthetic Ethnography: Field Devices for the Qualitative Study of Generative Models. *SocArXiv*. https://doi.org/10.31235/osf.io/zvew4

Dove, G., Halskov, K., Forlizzi, J. & Zimmerman, J. (2017). UX Design Innovation: Challenges for Working with Machine Learning as a Design Material. *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (pp. 278–288). ACM Press. https://dl.acm.org/doi/10.1145/3025453.3025739

Duguay, S., & Gold-Apel, H. (2023). Stumbling Blocks and Alternative Paths: Reconsidering the Walkthrough Method for Analyzing Apps. *Social Media + Society, 9*(1). https://doi.org/10.1177/20563051231158822

Dwivedi, Y. K., Kshetri, N., Hughes, L., Slade, E. L., Jeyaraj, A., Kar, A., Baabdullah, A.,M., Koohang, A., Raghavan, V., Ahuja,M., Albanna, H., Albashrawi, M.A., Al-Busaidi, A.S., Balakrishnan, J., Barlette, Y., Basu, S., Bose, I., Brooks, L., Buhalis, D., Carter, L., Chowdhury, S., Crick, T., Cunningham, S.W., Davies, G.H., Davison, R.M., Dé, R., Dennehy, D., Duan, Y., Dubey, R., Dwivedi, R., Edwards, J.S., Flavián, C., Gauld, R., Grover, V., Hu, M.-C., Janssen, M., Jones, P., Junglas, I., Khorana, S., Kraus, S., Larsen, K.R., Latreille, P., Laumer, S., Malik, F.T., Mardani, A., Mariani, M., Mithas, S., Mogaji, E., Nord, J.H., O'Connor, S., Okumus, F., Pagani, M., Pandey, N., Papagiannidis, S., Pappas, I.,O., Pathak, N., Pries-Heje, J., Raman, R., Rana, N.P., Rehm, S.-V., Ribeiro-Navarrete, S., Richter, A., Rowe, F., Sarker, S., Carsten Stahl, S., Kumar Tiwari, M., van der Aalst, W., Venkatesh, V., Viglia, G., Wade, M., Walton, P., Wirtz, J., & Wright, R. (2023). "So What if ChatGPT Wrote it?" Multidisciplinary Perspectives on Opportunities, Challenges and Implications of Generative Conversational AI for Research, Practice and Policy. *International Journal of Information Management*, *71*, 102642. https://doi.org/10.1016/j.ijinfomgt.2023.102642

Edkie, A., Pandey, D., & Roy, S. (2020). Murf.AI. https://murf.ai/

Farooq, M., Buzdar, H. Q. & Muhammad, S. (2023). AI-Enhanced Social Sciences: A Systematic Literature Review and Bibliographic Analysis of Web of Science Published Research Papers. *Pakistan Journal of Society, Education and Language (PJSEL)*, *10*(1), 250–267.

Ferrara, E. (2024). Fairness and Bias in Artificial Intelligence: A Brief Survey of Sources, Impacts, and Mitigation Strategies. *Sci*, *6*(1). https://doi.org/10.3390/sci6010003

Franklin, U. (1990). *The Real World of Technology*. Toronto: CBC.

García-Peñalvo, F., & Vázquez-Ingelmo, A. (2023). What Do We Mean by GenAI? A Systematic Mapping of The Evolution, Trends, and Techniques Involved in Generative AI. *International Journal of Interactive Multimedia and Artificial Intelligence*, *8*(4), 7. https://doi.org/10.9781/ijimai.2023.07.006

Gaspar, B. (2023). Cientistas divulgam 10 diretrizes para a Educação lidar com a Inteligência Artificial. Fepesp - Federação Dos Professores Do Estado de São Paulo. https://fepesp.org.br/noticia/cientistas-divulgam-10-diretrizes-para-a-educacao-lidar-com-a-inteligencia-artificial/

Google Creative Lab (2017). Teachable Machine. [software]. https://teachablemachine.withgoogle.com/

Gorska, A.M., & Jemielniak, D. (2023). The Invisible Women: Uncovering Gender Bias in AI-generated Images of Professionals. *Feminist Media Studies*, *23*(8), 4370–4375. https://doi.org/10.1080/14680777.2023.2263659

Goulart, J. (2024). Silvio Meira: 'Estamos na era da pedra lascada da IA, mas o futuro chega em 800 dias'. *Brazil Journal*, 16 March. https://braziljournal.com/silvio-meira-estamos-na-era-da-pedra-lascada-da-ia-mas-o-futuro-chega-em-800-dias/

Gray, J., Bounegru, L., Rogers, R., Venturini, T., Ricci, D., Meunier, A., Mauri, M., Niederer, S., Sánchez Querubín, N., Tuters, M., Kimbell, L., & Munk, K. (2022). Engaged Research-led Teaching: Composing Collective Inquiry with Digital Methods and Data. *Digital Culture & Education*, *14*(3), 55–86.https://www.digitalcultureandeducation.com/volume-14-3

Graziani, M., Dutkiewicz, L., Calvaresi, D., Amorim, J. P., Yordanova, K., Vered, M., Nair, R., Henriques Abreu, P., Blanke, T., Pulignano, V., Prior, J.O., Lauwaert, L., Reijers, W., Depeursinge, A., Andrearczyk, V., & Müller, H. (2023). A Global Taxonomy of Interpretable AI: Unifying the Terminology for the Technical and Social Sciences. *Artificial Intelligence Review*, *56*(4), 3473–3504. https://link.springer.com/article/10.1007/s10462-022-10256-8

Greene, C. (2023). AI and the Social Sciences: Why All Variables are Not Created Equal. *Res Publica*, *29*(2), 303–319. https://doi.org/10.1007/s11158-022-09544-5

Hartman, J., Schwenzow, J., & Witte, M. (2023). The Political Ideology of Conversational AI: Converging Evidence on ChatGPT's Pro-environmental, Left-libertarian Orientation. *arXiv*. https://doi.org/10.2139/ssrn.4316084

Hoel, A. S. (2012). Technics of Thinking. In A.S. Hoel & I. Folkvord (Eds.), *Ernst Cassirer on Form and Technology: Contemporary Readings* (pp. 65–91). London: Palgrave Macmillan.

Honig, C., Rios, S., & Oliveira, E. (2023). A Tool for Learning: Classroom Use-cases for Generative AI. *The Chemical Engineer*, 1 June. https://www.thechemicalengineer.com/features/a-tool-for-learning-classroom-use-cases-for-generative-ai/

Hooks, B. (1992). *Black Looks: Race and Representation*. Boston, MA: End Press.

Jacomy, M., Venturini, T., Heymann, S., & Bastian, M. (2014). ForceAtlas2, a Continuous Graph Layout Algorithm for Handy Network Visualization Designed for the Gephi Software. *PLoS ONE*, *9*(6), e98679. https://doi.org/10.1371/journal.pone.0098679

Kendall, A., Grimes, M., & Cipolla, R. (2016). PoseNet: A Convolutional Network for Real-Time 6-DOF Camera Relocalization. *arXiv*. https://doi.org/10.1109/ICCV.2015.336

Koplin, J.J. (2023). Dual-use Implications of AI text Generation. *Ethics and Information Technology*, *25*(2), 32. https://doi.org/10.1007/s10676-023-09703-z

Leshkevich, T., & Motozhanets, A. (2022). Social Perception of Artificial Intelligence and Digitization of Cultural Heritage: Russian Context. *Applied Sciences*, *12*(5), 2712. https://doi.org/10.3390/app12052712

Limewire. (2023). BlueWillow. [Software]. https://www.bluewillow.ai/

Luccioni, A.S., Akiki, C., Mitchell, M., & Jernite, Y. (2023). Stable Bias: Evaluating Societal Representations in Diffusion Models. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, & S. Levine (Eds.), *Advances in Neural Information Processing Systems* (pp. 56338–56351). New York, NY: Curran Associates.

Maier, N., Parodi, F., & Verna, S. (2004). DownThemAll! (v4.12.1). [Web browser plugin]. https://www.downthemall.org/

Manovich, L. (2013). Museum Without Walls, Art History Without Names: Methods and Concepts for Media Visualization. In C. Vernallis, A. Herzog & J. Richardson (Eds.), *The Oxford Handbook of Sound and Image in Digital Media* (pp. 252–278). Oxford: Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199757640.013.005

Manovich, L. (2020). *Cultural Analytics*. Cambridge, MA: MIT Press. https://doi.org/10.7551/mitpress/11214.001.0001

Marres, N. (2017). *Digital Sociology: The Reinvention of Social Research*. London: Wiley.

Mauri, M. & Ciuccarelli, P. (2016). Designing Diagrams for Social Issues. Future Focused Thinking - DRS International Conference 2016. https://doi.org/10.21606/drs.2016.185

Mauri, M., Briones, M.A., Gobbo, B. & Colombo, G. (2020). Research Protocol Diagrams as Didatic Tools to Act Critically in Dataset Design Processes. *INTED2020 Proceedings*, (pp. 9034–9043). https://doi.org/10.21125/inted.2020.2470

Microsoft. (2023). Bing Image Creator. https://www.bing.com/images/create

Midjourney Inc. (2022). Midjourney (Version 5.2). https://www.midjourney.com/

Mostaque, E. (2019). Stability.ai. https://stability.ai/

Nicoletti, L., & Bass, D. (2023). Humans Are Biased. Generative AI Is Even Worse. *Bloomberg*, 9 June. https://www.bloomberg.com/graphics/2023-generative-ai-bias/

Noble, U.S. (2013). Google Search: Hyper-visibility as a Means of Rendering Black Women and Girls Invisible. *In Visible Culture*, *19*. https://doi.org/10.47761/494a02f6.50883fff

Noble, S.U. (2018). *Algorithms of Oppression*. New York, NY: New York University Press. https://doi.org/10.18574/nyu/9781479833641.001.0001

Omena, J.J. (2021a). *Digital Methods and Technicity-of-the-Mediums. From Regimes of Functioning to Digital Research*. [Doctoral Dissertation, Universidade NOVA de Lisboa]. http://hdl.handle.net/10362/127961

Omena, J.J., Pilipets, E., Gobbo, B., & Chao, J. (2021b). The Potentials of Google Vision API-based Networks to Study Natively Digital Images. *Revista Diseña*, *19*. https://doi.org/10.7764/disena.19.article.1

Omena, J.J. (2022). Technicity-of-the-mediums. In A. Ceron (Ed.), *Elgar Encyclopedia of Technology and Politics* (pp. 77–81). Cheltenham: Elgar.

OpenAI. (2023). ChatGPT. *Large Language Model*. OPenAI. https://chat.openai.com

OpenAI, Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Leoni Aleman, F., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., Avila, R., Babuschkin, I., Balaji, S., Balcom, V., Baltescu, P., Bao, H., Bavarian, M., Belgum, J., Bello, I., Berdine, J., Bernadett-Shapiro, G., Berner, C., Bogdonoff, L., Boiko, O., Boyd, M., Brakman, A.L., Brockman, G., Brooks, T., Brundage, M., Button, K., Cai, T., Campbell, R., Cann, A., Carey, B., Carlson, C., Carmichael, R., Chan, R., Chang, C., Chantzis, F., Chen, D., Chen, S., Chen, R., Chen, J., Chen, M., Chess, B., Cho, C., Chu, C., Won Chung, H., Cummings, D., Currier, J., Dai, Y., Decareaux, C., Degry, T., Deutsch, N., Deville, D., Dhar, A., Dohan, D., Dowling, S., Dunning, S., Ecoffet, A., Eleti, A., Eloundou, T., Farhi, D., Fedus, L., Felix, N., Posada Fishman, S., Forte, J., Fulford, I., Gao, L., Georges, E., Gibson, C., Goel, V., Gogineni, T., Goh, G., Gontijo-Lopes, R., Gordon, J., Grafstein, M., Gray, S., Greene, R., Gross, J., Shane Gu, S., Guo, Y., Hallacy, C., Han, J., Harris, J., He, Y., Heaton, M., Heidecke, J., Hesse, C., Hickey, A., Hickey, W., Hoeschele, P., Houghton, B., Hsu, K., Hu, S., Hu, X., Huizinga, J., Jain, S., Jain, S., Jang, J., Jiang, A., Jiang, R., Jin, H., Jin, D., Jomoto, S., Jonn, B., Jun, H., Kaftan, T., Kaiser, Ł., Kamali, A., Kanitscheider, I., Shirish Keskar, N., Khan, T., Kilpatrick, L., Wook Kim, J., Kim, C., Kim, Y., Kirchner, J.H., Kiros, J., Knight, M., Kokotajlo, D., Kondraciuk, Ł., Kondrich, A., Konstantinidis, Kosic, K., Krueger, G., Kuo, V., Lampe, M., Lan, I., Lee, T., Leike, J., Leung, J., Levy, A.D., Ming Li, C., Lim, R., Lin, M., Lin, S., Litwin, M., Lopez, T., Lowe, R., Lue, P., Makanju, A., Malfacini, K., Manning, S., Markov, T., Markovski,Y., Martin, B., Mayer, K., Mayne, A., McGrew, B., Mayer McKinney, S., McLeavey, C., McMillan, P., McNeil, J., Medina, D., Mehta, A., Menick, J., Metz, L., Mishchenko, A., Mishkin, P., Monaco, V., Morikawa, E., Mossing, D., Mu, T., Murati, M., Murk, O., Mély, D., Nair, A., Nakano, R., Nayak, R., Neelakantan, A., Ngo, R., Noh, H., Ouyang, L., O'Keefe, C., Pachocki, J.,Paino, A., Palermo, J., Pantuliano, A., Parascandolo, G., Parish, J., Parparita, E., Passos, A., Pavlov, M., Peng, A., Perelman, A., de Avila Belbute Peres, F., Petrov, M., Ponde de Oliveira Pinto, H., Rai Pokorny, .M., Pokrass, M., Pong, V.,H., Powell, T., Power, A., Power, B., Proehl, E., Puri, R., Radford, A., Rae, J., Ramesh, A., Cameron Raymond, Real, F., Rimbach, K., Ross, C., Rotsted, B., Roussez, H., Ryder, N., Saltarelli, M., Sanders, T., Santurkar, S., Sastry, G., Schmidt, H., Schnurr, D., Schulman, J., Selsam, D., Sheppard, K., Sherbakov, T., Shieh, J., Shoker, S., Shyam, P., Sidor, S., Sigler, E., Simens, M., Sitkin, J., Slama, K., Sohl, I., Sokolowsky, B., Song, Y., Staudacher, N., Such, F.P., Summers, N., Sutskever, I., Tang, J., Tezak, N., Thompson, M.B., Tillet, P., Tootoonchian, A., Tseng, E., Tuggle, P., Turley, N., Tworek, J., Cerón Uribe, F.J., Vallone, A., Vijayvergiya, A., Voss, C., Wainwright, C., Wang, J.J., Wang, A., Wang, B., Ward, J., Wei, J., Weinmann, C.J., Welihinda, A., Welinder, P., Weng, J., Weng, L., Wiethoff, M., Willner, D., Winter, C., Wolrich, S., Wong, H., Workman, L., Wu, S., Wu, J., Wu, M., Xiao, K., Xu, T., Yoo, S., Yu, K., Yuan, Q., Zaremba, W., Zellers, R., Zhang, C., Zhang, M., Zhao, S., Zheng, T., Zhuang, J., Zhuk, W., & Zoph, B. (2024). GPT-4 Technical Report. *arXiv*. http://arxiv.org/abs/2303.08774

Pask, G. (1975). Minds and Media in Education and Entertainment: Some Theoretical Comments Illustrated by the Design and Operation of a System for Exteriorizing and Manipulating Individual Theses. In R. Trappl & G. Pask (Eds.), *Progress in Cybernetics and System Research* (pp. 38–50). London: Hemisphere.

Peeters, S. (2023). Zeeschuimer (Version 1.4). [Firefox plugin]. https://doi.org/10.5117/CCR2022.2.007.HAGE

Perez, J., Castro, M., & Lopez, G. (2023). Serious Games and AI: Challenges and Opportunities for Computational Social Science. *IEEE Access*, *11*, 62051–62061. https://doi.org/10.1109/ACCESS.2023.3286695

Peeters, S., & Hagen, S. (2022). The 4CAT Capture and Analysis Toolkit: A Modular Tool for Transparent and Traceable Social Media Research. *Computational Communication Research*, *4*(2), 571–589. https://computationalcommunication.org/ccr/article/view/120

Popescu, A., & Schut, A. (2023). Generative AI in Creative Design Processes: aDive into Possible Cognitive Biases. In D. De Sainz Molestina, L. Galluzzo, F. Rizzo & D. Spallazzo (Eds.), *IASDR 2023: Life-Changing Design* (pp. 1–10). https://doi.org/10.21606/iasdr.2023.784

Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., & Chen, M. (2022). Hierarchical Text-Conditional Image Generation with CLIP Latents. *arXiv*. http://arxiv.org/abs/2204.06125

Rieder, B. (2020). *Engines of Order: a Mechanology of Algorithmic Techniques*. Amsterdam: Amsterdam University Press.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P. & Ommer, B. (2021). High-Resolution Image Synthesis with Latent Diffusion Models. *arXiv*. https://doi.org/10.1109/CVPR52688.2022.01042

Rogers, R. (2013). *Digital Methods*. Cambridge, MA: MIT Press.

Rogers, R. & Lewthwaite, S. (2019). Teaching Digital Methods: Interview. with Richard Rogers. *Revista Diseña*, *14*, 12–37. https://doi.org/10.7764/disena.14.12-37

Rogers, S., & Cairo, A. (2022). TwoTone. https://twotone.io/

Rozado, D. (2023). The Political Biases of ChatGPT. *Social Sciences*, *12*(3), 148. https://doi.org/10.3390/socsci12030148

Russell Group. (2023). New Principles on Use of AI in Education. *The Russell Group*, 4 June. https://russellgroup.ac.uk/news/new-principles-on-use-of-ai-in-education/

Salvaggio, E. (2022). How to Read an AI Image. *Cybernetic Forests*, 2 October. https://www.cyberneticforests.com/news/how-to-read-an-ai-image

Shameem, S. (2022). Lexica AI. https://lexica.art/

Sharma, S. (2024). *Understanding Digital Racism: Networks, Algorithms, Scale*. Lanham, MD:Rowman & Littlefield.

Shrestha, Y.R., von Krogh, G., & Feuerriegel, S. (2023). Building Open-Source AI. *Nature Computational Science*, *3*, 908–911 http://dx.doi.org/10.2139/ssrn.4614280

Silva, T. (2023). *Mapeamento de Danos e Discriminação Algorítmica*. Desvelar. https://desvelar.org/casos-de-discriminacao-algoritmica/

Sinclair, D., Dowdeswell, T., & Goltz, N. (2023). Artificially Intelligent Sex Bots and Female Slavery: Social Science and Jewish Legal and Ethical Perspectives. *Information & Communications Technology Law*, *32*(3), 328–355. https://doi.org/10.1080/13600834.2022.2154050

Sinclair, S., & Rockwell, G. (2003). Voyant Tools (v2.6.13). [Software]. https://voyant-tools.org/

Souza, R. [@renatasouzario]. (2023). Racismo nas plataformas de inteligência artificial! [Video]. *Instagram*, 25 October. https://www.instagram.com/reel/Cy1p6EQpwXB/?igshid=MzRlODBiNWFlZA%3D%3D

Stokel-Walker, C., & Van Noorden, R. (2023). What ChatGPT and Generative AI Mean for Science. *Nature*, *614*(7947), 214–216. https://doi.org/10.1038/d41586-023-00340-6

Sun, L., Wei, M., Sun, Y., Suh, Y. J., Shen, L., & Yang, S. (2023). Smiling Women Pitching Down: Auditing Representational and Presentational Gender Biases in Image Generative AI. *arXiv*. https://doi.org/10.1093/jcmc/zmad045

The DigiKam Team. (2001). digiKam (v8.3.0). [Software]. https://www.digikam.org/

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., Bikel, D., Blecher, L., Canton Ferrer, C., Chen, M., Cucurull, G., Esiobu, D., Fernandes, J., Fu, J., Fu, W., Fuller, B., Gao, C., Goswami, V., Goyal, N., Hartshorn, A., Hosseini, S., Hou, R., Inan, H., Kardas, M., Kerkez, V., Khabsa, M., Kloumann, I., Korenev, A., Singh Koura, P., Lachaux, M.-A., Lavril, T., Lee, J., Liskovich, D., Lu, Y., Mao, Y., Martinet, X., Mihaylov, T., Mishra, P., Molybog, M., Nie, Y., Poulton, A., Reizenstein, J., Rungta, R., Saladi, K., Schelten, A., Silva, R., Smith, E.M., Subramanian, R., Tan, X.E., Tang, B., Taylor, R., Williams, A., Kuan, J.X., Xu, P., Yan, Z., Zarov, I., Zhang, Y., Fan, A., Kambadur, M., Narang, S., Rodriguez, A., Stojnic, R., Edunov, S., & Scialom, T. (2023). Llama 2: Open Foundation and Fine-Tuned Chat Models. *arXiv*. http://arxiv.org/abs/2307.09288

Valenzuela, C., Matamala, A., & Germanidis, A. (2018). RunwayML. https://runwayml.com/

Visual Computing Group. (2018). Image Sorter (v4). [Software]. https://visual-computing.com/project/imagesorter/

Visual Crossing Corporation. (2003). Visual Crossing. https://www.visualcrossing.com/

Vogel, K.M. (2021). Big Data, AI, Platforms, and the Future of the U.S. Intelligence Workforce: A Research Agenda. *IEEE Technology and Society Magazine*, *40*(3), 84–92. https://doi.org/10.1109/MTS.2021.3104384

Wang, F.-Y., Ding, W., Wang, X., Garibaldi, J., Teng, S., Imre, R., & Olaverri-Monreal, C. (2022). The DAO to DeSci: AI for Free, Fair, and Responsibility Sensitive Sciences. *IEEE Intelligent Systems*, *37*(2), 16–22. https://doi.org/10.1109/MIS.2022.3167070

Wojcicki, A. (2020). Artflow AI. https://app.artflow.ai

Yu, C., Tschanz-Egger, J.L., & Souto, M. (2023). Tomato Girl Summer. *Designing With: A New Educational Module to Integrate Artificial Intelligence, Machine Learning and Data Visualization in Design Curricula*, 19 June. https://master-interaction-design.notion.site/Tomato-Girl-Summer-07dcf86e607e44d5b00b5d8cd9524a75

Zajko, M. (2021). Conservative AI and Social Inequality: Conceptualizing Alternatives to Bias through Social Theory. *AI and Society*, *36*(3), 1047–1056. https://doi.org/10.1007/s00146-021-01153-9

Janna Joceli Omena – Department of Digital Humanities, King's College London (United Kingdom)
https://orcid.org/0000-0001-8445-9502 | J.J.Omena@kcl.ac.uk
https://thesocialplatforms.wordpress.com/
Janna Joceli Omena is a Lecturer (Assistant Professor) in Digital Methods at the Department of Digital Humanities, King's College London (UK). As a digital methodologist, her work is rooted in understanding the epistemological dimension of AI, web technologies, digital objects, and research software in method design and implementation and their role in research and knowledge production practices. Her research interest covers the practice and theory of digital methods, where she investigates the technicity of computational media from conceptual, technical, and empirical perspectives, using this knowledge to develop reproducible methodologies and research software.

Antonella Autuori – University of Applied Sciences and Arts of Southern Switzerland (SUPSI) (Switzerland); RMIT University, Melbourne (Australia)
https://orcid.org/0000-0002-5725-8446
https://www.supsi.ch/antonella-autuori
Antonella Autuori is a Design Researcher working at the Institute of Design at the University of Applied Sciences of Southern Switzerland (SUPSI). Her research focuses on defining new methodologies at the intersection of artificial intelligence and design, with a particular emphasis on the educational dimension. She is also a PhD student in Design at RMIT in Melbourne (Australia), where she is investigating how design can effectively mitigate the impact of biases inherent in the representation of reality by generative AI, and how it can promote a culture of reflective practice in the use of AI technologies.

Eduardo Leite Vasconcelos – Universidade Federal da Bahia (UFBA) (Brazil)
https://orcid.org/0000-0002-0937-395X
https://www.researchgate.net/profile/Eduardo-Vasconcelos-6
Eduardo Leite Vasconcelos is a digital journalism and imagery, photography, photojournalism, and visual methodologies Researcher. He holds a PhD in Contemporary Communication and Culture from Universidade Federal da Bahia, in Brazil, and is a member of the Online Journalism Research Group (GJOL/UFBA).

Matteo Subet – University of Applied Sciences and Arts of Southern Switzerland (SUPSI)
https://orcid.org/0009-0003-4769-9568
https://link.zumat.ch/
Matteo Subet is a Teaching and Research Assistant at the SUPSI Design Institute and the MA in Interaction Design (Switzerland). His research interests are human-computer interaction, design education and AI-driven systems in the design field, with a focus on new technologies and physical computing.

Massimo Botta – University of Applied Sciences and Arts of Southern Switzerland (SUPSI)
https://orcid.org/0009-0005-8894-2661
https://www.supsi.ch/massimo-botta
Massimo Botta PhD, Professor, is the Head of the MA in Interaction Design and Head of the Research Area in Interaction Design at the Institute of Design of SUPSI (Switzerland). His research interests include design education and methodologies, information visualization and human-computer interaction. He plays an active role in promoting the international debate on the evolution of design research with other sciences and the development of the third cycle of education (Ph.D.) in design through the organization of international conferences and the editing of publications.

Symposium: Repurposing Generative AI for Social Research – *peer-reviewed*
Edited by Federico Pilati, Anders Kristian Munk, and Tommaso Venturini
https://doi.org/10.6092/issn.1971-8853/19576

Sociologica. V.18 N.2 (2024)
ISSN 1971-8853
https://sociologica.unibo.it/

The Problems of LLM-generated Data in Social Science Research

Luca Rossi* [a] Katherine Harrison [c] Irina Shklovski [b, c]

[a] Department of Digital Design, IT University of Copenhagen (Denmark)
[b] Department of Computer Science, Department of Communication, University of Copenhagen (Denmark)
[c] Department of Thematic Studies – Gender Studies, Linköping University (Sweden)

Submitted: May 17, 2024 – Revised version: September 6, 2024
Accepted: September 23, 2024 – Published: October 30, 2024

Abstract

Beyond being used as fast and cheap annotators for otherwise complex classification tasks, LLMs have seen a growing adoption for generating synthetic data for social science and design research. Researchers have used LLM-generated data for data augmentation and prototyping, as well as for direct analysis where LLMs acted as proxies for real human subjects. LLM-based synthetic data build on fundamentally different epistemological assumptions than previous synthetically generated data and are justified by a different set of considerations. In this essay, we explore the various ways in which LLMs have been used to generate research data and consider the underlying epistemological (and accompanying methodological) assumptions. We challenge some of the assumptions made about LLM-generated data, and we highlight the main challenges that social sciences and humanities need to address if they want to adopt LLMs as synthetic data generators.

Keywords: LLM; synthetic data; social science; research methods.

Acknowledgements

This work was partially supported by the Wallenberg AI, Autonomous Systems and Software Program – Humanity and Society (WASP-HS) funded by the Marianne and Marcus Wallenberg Foundation.

* lucr@itu.dk

1 Introduction

Despite the novelty of the technology, Large Language Models (LLMs) have seen nascent adoption by social scientists for research purposes in part due to the widespread public availability of such tools made possible by commercial offers. Within the social sciences we observe two approaches to LLM-produced textual content that seem to dominate. First, there are many studies that have creatively applied social science research methods to study algorithmic systems themselves (for an overview see Moats & Seaver, 2019; and articles in this Symposium). In particular, with the advent of LLMs we see a range of applications of social science methods from psychology experiments (Almeida et al., 2024) to variations on ethnography (Demuro & Gurney, 2024) as alternative ways to understand the capacities and limits of these technologies. Second, there are a number of studies that have used LLMs to augment existing social science methods either in data analysis or data generation (Møller et al., 2024). While all of these deserve discussion, in this essay we focus on the use of LLMs to produce — or experiment with producing — synthetic data as research material for social science questions. In particular, we are concerned with studies that seek to use such LLM-generated data to model or make inferences about how people might act, react, or respond in a variety of situations.

In a widely cited article, Grossmann et al. (2023) argue that the capacity of LLMs for "simulating human-like responses and behaviors offers opportunities to test theories and hypotheses about human behavior at great scale and speed" (p. 1108). The authors go as far as to imagine that "LLMs may supplant human participants for data collection" (p. 1109). This enthusiasm seems to be shared by an increasing number of scholars (e.g., Jansen et al., 2023; Aher et al., 2023; Argyle et al., 2023; Hämäläinen et al., 2023; Törnberg et al., 2023). At the same time, many scholars question some of the assumptions underlying these studies and suggest caution in deploying LLMs as social science data-generation mechanisms (e.g., von der Heyde et al., 2023; Bisbee et al., 2023; Agnew et al., 2024).

If LLMs could generate data as if they were "participants", answering questions, making decisions or arguing for those, this would certainly represent a seismic shift for many social sciences that have often been struggling with hard-to-find participants, expensive data collection and questionable convenience samples. Yet while Grossmann et al. (2023) might claim that LLMs are now able to produce language in a "contextually aware and semantically accurate fashion" (p. 1108) current evidence within NLP research suggests that this is not quite the case just yet (Cui et al., 2023). While LLMs might represent a tempting opportunity for social science research given the ease and apparent facility in language production in response to carefully worded prompts, the use of such data for generating insights about people requires serious critical consideration.

In what follows, we review the extant literature and the emergent debate on the topic and discuss what we see as the main concerns with the use of LLMs as a source of social science research data.

First, we will discuss what type of synthetic data is actually produced by LLMs and in what way it differs from other approaches to producing synthetic data. Second, we will explore the implications of the cautions Grossmann et al. (2023) note — the fidelity of training data and considerations of its representativeness, the problem of LLM biases, the challenges of transformer-type models and their propensity for hallucinations, the emergent art of prompt engineering and the idea of benchmark selection. Third, we will consider what kind of knowledge can be gained from the analysis of LLM-produced data if we were to imagine that the problems of benchmarking, transparency in model training, and data fidelity can be addressed and what sort of legitimacy this knowledge might claim. Ultimately, we reflect on the proposition recently put forward by Bail (2024), that researchers could work towards the creation of open-source LLMs specifically trained and deployed for social research, by highlighting what we see as the major challenges that would need to be addressed.

2 Particularities of LLM-generated Data

LLM-produced data can be seen as a type of generated synthetic data. Synthetic data are datasets generated using purpose-built mathematical models or algorithms (Jordon et al., 2022) instead of being extracted from existing digital systems or produced through particular forms of data collection. The idea of synthetic data has a significant history in statistics and particularly within governmental and public institutions that need to make population data available for analysis while ensuring confidentiality (Abowd & Vilhuber, 2008) where population datasets have typically been treated through data augmentation and statistical disclosure control techniques (Raghunathan, 2021). Synthetic simulation data also has a long history in areas such as computer vision, which typically uses model specifications to generate datasets (Nikolenko, 2021). The advent of greater computational capacities and more complex machine learning models made synthetic data augmentation and generation faster, cheaper, less complex, and increasingly popular (Jordon et al., 2022).

Typically, synthetic data are seen to address any or all of three primary data challenges: data scarcity, data privacy, and data bias (Van der Schaar & Qian, 2023). While there is no doubt that generating synthetic data addresses the issue of scarcity by producing more data given a set of specifications, whether these data are of sufficient quality to be useful and whether they are able to address the challenges of privacy and bias depends on context (Abowd & Vilhuber, 2008; Belgodere et al., 2023; Figueira & Vaz, 2022; Jacobsen, 2023). With the advent of generative AI models in the development of synthetic data pipelines, we see the development of new frameworks proposed to evaluate the quality of the generated datasets. For example, Eigenschink et al. (2023) propose five assessment criteria for synthetically produced datasets: representativeness, novelty, realism, diversity, and coherence. They argue that high-quality synthetic data produced via generative AI should be able to capture population-level properties of the original data, create novel data points that are realistic (given what we know of the original data), and show internal diversity while maintaining coherency with the original data. They also note that the importance of the individual criteria varies significantly across domains and that the ways in which such criteria should be tested differ. Without delving too deeply into the specific framework, it is clear that the focus is on the ability of the synthetic data to reproduce key characteristics of the original data.

The development and rapid adoption of LLMs have led to the use of these systems for a broad range of data generation tasks. Whitney & Norman (2024) distinguish between generated, augmented, and procedurally created synthetic datasets differentiating between them "based on how derivative of a real-world training dataset they are" (p. 4). Procedurally created synthetic data rely on purpose-built models that create data along a set of explicitly pre-specified parameters, while generated data, such as what LLMs produce, arise from a model abstracting from a training dataset in response to a particular input. Here, the training dataset is crucial for ensuring that the resulting dataset is usefully similar to the data we might expect to collect from people. LLM-driven systems, however, are not built with faithful data generation as a goal. Rather than aiming to produce data that resemble a given original dataset, LLMs have been trained to predict the occurrence of the next most likely letter, word, or group of words

https://doi.org/10.6092/issn.1971-8853/19576

given the linguistic patterns of a text. Trained on ever larger amounts of data, LLMs are able to mimic human-produced content (Jakesch et al., 2023) showing emergent abilities to perform tasks that were not explicitly present in the training data[1] (Radford et al., 2019; Wei et al., 2022). There is no effort here to adhere to particular pre-existing patterns in the data.

Given the ease with which current LLM implementations can generate human-like text on a near-infinite number of topics, it was not a huge leap to imagine the use of these technologies for generating data for research purposes. Seen as a convenient type of data, such LLM-generated datasets are typically obtained given sufficient effort in prompt engineering and, occasionally, the use of different available LLM implementations for comparison (Dillion et al., 2023, Horton, 2023). It is possible to classify LLMs as a type of synthetic data generator (Jordon et al., 2022). Keeping in mind that the goal of synthetic data generators is to produce data that resemble particular aspects of real data while addressing issues such as privacy, lack of diversity or data scarcity, one might ask if these goals are achieved by LLMs.

On the surface, LLMs can certainly generate data that mimics human-produced data, thus potentially resembling aspects of real data. Yet there is one important caveat to consider. Synthetic data is typically evaluated for utility and fidelity — how useful they are for a particular task and how well they resemble a real dataset given parameters important to the task. Here different types of fidelity may be considered but the important question is what is necessary for the task, but they require some capacity to compare the data we intend to mimic or augment and the synthetic output. At times, fidelity issues can be quite insidious, as Johnson & Hajisharif (2024) demonstrate a lack of intersectional fidelity in their tests with GAN-produced synthetic census data. How are we to evaluate the fidelity of an LLM-produced dataset, especially given the lack of access to the models generating the data or their training data? How might we evaluate differences in these data and their implications? Despite continuous advances, LLMs continue to display a lack of internal consistency in output. For example, Atil et al. (2024) have recently reported how, when systematically studied, LLMs show a lack of stability even when the input is the same. This was observed despite all the hyper-parameters being set to maximize the deterministic nature of the model. Importantly, stability not only varied between different models (both commercial and open source) but also within the same model as a function of the task.

Questions of utility and fidelity, of course, will vary depending on the particular applications of LLM-produced data, which we will address in the rest of the essay.

3 Applications of LLM-produced Data in Social Science Research

While clear-cut divisions are inherently problematic, here we identify what we see as major streams in LLM adoption for data production in the context of different types of social science research. Organizing the existing methodological experimentation around three directions allows us to see the existing similarities and differences between these approaches as well as the shared underlying assumptions about "what" LLMs can do and why. It is worth noting that while LLMs have been used to generate a variety of data we will generally treat this as textual data for two main reasons. First, the underlying training data of LLMs are essentially textual in nature. Second, even when LLMs are used to produce numeric values (e.g., producing a

1. Several researchers have questioned the concept of emergent capabilities (see Schaeffer et al., 2024). The actual nature of these capabilities as well as their origin is irrelevant for the point of the specific article.

number that expresses agreement or disagreement on a scale) that is achieved through a textual prompt that asks the model to express the output as a number that would, otherwise, be textual.

When researchers use LLMs to produce synthetic research data, they are leveraging the underlying large amount of training data that characterizes these models as a proxy for individual data or data about specific groups and populations. This underlying, often unspoken, premise holds regardless of the specific research design and the specific LLM used. Within this perspective, there is no difference between using GPT-4, LLAMA-3 or Claude 3.5. What appears to be revolutionary is that LLM technology has reached a scale that allows emergent and unprecedented capabilities (Grossmann et al., 2023), irrespective of specific implementations. Nevertheless, the vast majority of research in this overview relies on OpenAI's GPT-3.5 and GPT-4 models. This is not due to any explicit analysis of the specific capabilities of GPT versus alternative models but rather due to the easy access provided by OpenAI's well-developed set of APIs.

3.1 LLMs as an Improved Version of Agent-Based Modeling (ABM)

The excitement about LLM-produced data hinges on the model's capacity to simulate human-produced text (Grossmann et al., 2023; Bail, 2024). Controlled through carefully developed prompts, such text production has been used to augment or even replace other forms of agent-based simulation by several scholars (Park et al., 2023; Törnberg et al., 2023; Horton, 2023). While agent-based modeling has a long history, scholars in this domain have long struggled to overcome the limitations of ABM approaches: the necessary abstraction and simplification of the modeled context and the lack of capacity of these models to capture human discourse (Törnberg et al., 2023). Given LLMs' ability to produce text that reflects realistic human reasoning, such simulations would seem to address these major shortcomings of ABM.

In this context, LLMs have been used to create "personas" (Törnberg et al., 2023) through specific prompts defining the relevant personality traits for each manifestation. These agents were then used to create role-play situations following the prompted guidelines. The result showed a higher level of emergent behavior when compared to mechanistic ABMs (Tornberg et al., 2023). Park et al. (2023) have created an architecture based on ChatGPT to generate computational software agents that present what they term "believable simulations of human behavior." Such social simulacra (Park et al., 2022) are used to explore real-world scenarios with increased nuance, not available to more traditional ABM approaches (Wu et al., 2023), achieved by ensuring backward and forward continuity (Argyle et al., 2023) as well as extended memory (Park et al., 2023).

While the development of ABMs can be quite complex, authors point out that it is possible to generate autonomous goal-oriented agents by using LLMs with well-designed prompts quickly and at little cost (Phelps & Russel, 2023). This direction has generated a great amount of enthusiasm and has led to the creation of ad-hoc solutions where less technical researchers can deploy LLM-based social simulations (Rossetti et al., 2024).

Despite the excitement, there are some cautions in deploying these approaches. While some researchers find the results of such explorations convincing (Törnberg et al., 2023) or believable (Park et al., 2023), others note that there are limitations to how well such models are able to replicate human behavior in simulations of well-known contexts such as the iterated Prisoner's Dilemma (Phelps & Russel, 2023).

Nevertheless, simulations produced through ABM or LLM-based efforts do not need to be entirely faithful to particulars of human behavior. After all, following George Box's famous maxim, models can be useful even though all of them are wrong (Box, 1976). Where implementations of LLMs for modeling social contexts are used for insights into how people might act in a variety of situations, some additional caution is warranted. It is precisely because LLMs generate text, we notice an interesting trend towards personification (Jones et al., 2023) of these systems in the interpretation of results. In their exploration of how people relate to GPT-3 output, Jones et al. (2023) note that personification seems a common response, defining this as the tendency to seek a human-like intentionality behind the output. For example, Törnberg et al. (2023) present an interesting effort to simulate how different designs of social media platforms might affect the resulting toxicity of the posted content. They use LLMs to generate personas that then interact by producing simulated messages given prompts. They measure the toxicity of the resulting text and suggest particular designs as potentially more successful. However, in their interpretation, they seem to personify the simulated agents by noting that the agents were "[...] responding to the posts from the other side that trigger or upset them" (Törnberg et al., 2023, p. 6). Of course, LLMs can not be triggered or become upset, regardless of whether these systems are simulating a persona or simply producing text in response to a prompt. LLMs, after all, don't have emotions but such personification is common in interaction with systems that produce text (Jones et al., 2023). Törnberg et al. (2023) seem to rely on such personification as a causal explanation for the evidence of increasing toxicity in the simulation potentially over-interpreting or oversimplifying the implications of their data.

Interpretation, of course, is the linchpin of any social science research and the question remains how to interpret LLM-generated output in this context. ABM researchers readily admit the limits and oversimplifications of their models (Phelps & Russel, 2023). Yet as Box (1976) reminds us, knowing how and why our models are wrong is what enables us to make them useful. There is no doubt that LLM-generated output is "wrong" in the Box sense, but how and why are difficult questions to answer. Thus interpretation of results may rely on personification and naive comparison to the researcher's prior knowledge of contexts under study without any real relationship to what the output actually represents.

3.2 LLMs as Humans in the Bottle

A second stream of research uses LLMs to substitute research participants in what would traditionally be an experimental setting (Horton, 2023; Breum at al., 2023; Dillion et al., 2023). In this context, the role-playing ability of LLMs together with the ability to act according to specific instructions are used to generate particular interactions often between two instances of the same model. For example, treating LLMs as "implicit computation models of humans," Horton (2023) draws on classical experiments in behavioral economics to demonstrate how the use of LLMs to simulate socio-economic decision-making and outcomes can move beyond theoretical economics as a way to generate insights that could be tested using more expensive methods of research with people. Horton readily acknowledges that LLMs are just as wrong as mathematical models of economic behavior but demonstrates how they can provide useful insight. This approach has also been used to study whether LLMs can reproduce dynamics of persuasion typical of human social systems (Breum et al., 2023) or if they can replicate well-known economic and social psychological behaviors (Aher et al., 2023). Where some of this research turns social science methods to explore the limits and possibilities of LLMs, these studies also explore the potential of such approaches for advancing social science research in general.

One question this research explores is whether LLMs can "faithfully" reproduce human dynamics through text production (Aher et al., 2023). Some researchers focus on comparing the output of LLMs with the results of well-known psychological or economic experiments. This research attempts to make an argument for exactly how well such "implicit computational models of humans" (Horton, 2023, p. 2) can perform, in order to assess how reliable these models might be for new experimental efforts (Aher et al., 2023). As such, the criteria used to evaluate the resulting data quality — and ultimately the ability of the model to act "as a proxy for a human subject" — are largely based on the ability of the model to reproduce outcomes in well-known, previously published papers. For example, Horton (2023) uses LLMs to simulate outcomes of a decision-making scenario of allocating the federal budget between highway and car safety programs, originally presented in a well-known paper by Samuelsen and Zekhauser (1988). Results appear to show that the more advanced GPT-3 Davinci model can replicate the status quo bias demonstrated in the original paper.

Aher et al. (2023) propose the term "Turing Experiments (TE)" as a means of evaluating AI systems "in terms of its use in simulating human behavior in the context of a specific experiment" (p. 1). They replicate, among others, the famous controversial shock experiment designed by Milgram in 1963, where subjects were asked to shock the victim (an actor in another room) with an increasingly high voltage. The experiment was originally intended to demonstrate how far people are willing to go to conform to authority demands in the face of causing harm and pain to someone. While ideally, simulations, such as those presented by Horton (2023) or Aher et al. (2023), ought to be zero-shot, the fact that LLMs have been trained on the vast corpora of Internet data generally means that these data are likely to include prior descriptions of these famous experiments. To mitigate this factor, Aher et al. (2023) augmented the original experiment in ways that they argued maintained the integrity of the results. They compare the level of compliance observed by Milgram and reported in the 1963 publication with the level of "compliance" simulated by the LLM, noting the similarity between simulated and experimentally observed outcomes.

The idea here is that the similarity of LLMs' outcomes to published experimental data demonstrates how faithfully a model is capable of reproducing human behavior. This provides a legitimate argument for the use of these systems for validating new hypotheses about human behavior, especially where more traditional modes of data collection can be difficult or prohibitively expensive. Part of the problem with this argument is the fundamental assumption that prior experimental results are representative of human responses — an assumption that has been repeatedly called into question, especially around classic social psychology experiments of conformity conducted by Milgram & Ash (Greenwood, 2018; Henrich et al., 2010). The capacity of these models to reproduce such experiments is likely more reflective of a collective Western conviction that these experiments represent human behavior, rather than reflecting or representing human behavior. The famous psychology experiments were intended to demonstrate that our own beliefs and stories about why we do what we do are faulty. The question then is how should LLMs' output be interpreted correctly in such studies.

3.3 LLMs as Respondents to Surveys or Interviews

While this is the least common of the three streams of research and experimentation that we have identified, it is also the most problematic. The use of LLMs in assisting survey research spans the gamut from generating survey questions, pre-testing survey instruments, or analyzing data and summarizing findings (Jansen et al., 2023). Some research, however, has explored the potential of LLMs to generate data that is then analyzed. In this section, we share examples of how such data has been tested for predicting human responses in fields as varied as political theory, market research and design.

Although there are no studies yet that attempt to use LLM-produced data to make strong claims about human responses, several researchers are exploring this possibility. Argyle et al. (2023) presented one of the first efforts to demonstrate that LLMs can be used to generate data that replicates known distributions of particular response patterns in what they, similar to Horton's (2023) "homo silicus", call "silicone samples". They make the assumption that LLM output is based on underlying "human-like concept associations" where, "given basic human demographic background information" the output can model "underlying patterns between concepts, ideas, and attitudes that mirror those recorded from humans with matching backgrounds" (Argyle et al., 2023, p. 4). While such a statement is in agreement with the sentiment voiced by Grossmann et al. (2023), recent NLP research demonstrates that this is an overstatement of current LLM capacities. For example, transformer-based language models continue to have trouble generalizing beyond common linguistic constructions (Cui et al., 2023).

Similarly, Brand et al. (2023) explore the capacity of LLMs to respond to survey questions in a way that is consistent with economic theories and known consumer behavior patterns. Motoki et al. (2024) deploy several well-known survey instruments about organizational behavior and compare LLM-generated responses to published papers, noting that despite some limitations the outcomes do replicate human behavior and can potentially be used to validate survey instruments. In contrast, von der Heyde et al. (2023) generate LLM-based personas based on German voting data and show that the LLM-generated outcomes tend to be more biased and inaccurately predict voter choices. There is an emergent debate in the field where several studies have demonstrated that LLM-generated data tends to be significantly unrepresentative, arguing that perhaps such models are unfit for research applications (Simmons & Savinov, 2024; Bisbee et al., 2023; Santurkar et al., 2023).

Going beyond survey responses, Hämäläinen et al. (2023) explore whether LLMs can be productively used for qualitative research, specifically in design and user-experience research. They generate interview responses using persona-based prompts and compare the outcomes to published interview data. While there is an agreement that LLMs may not be particularly useful for predicting human responses to a range of cues, it is argued that such social simulations may nevertheless be useful for design purposes (Park et al., 2023; Hämäläinen et al., 2023). Designers have long used techniques such as developing personas and imagining responses to particular interactions with technology (Salminen et al., 2022) loosely based on research with potential users. Thus it is not a far cry to imagine how LLMs could be used for a similar purpose. Further, design research has frequently struggled with the problems of representation and inclusiveness — where user research focused on easily accessible people thus failing to address edge users and lacking diversity in samples (Sin et al., 2021; Elsayed-Ali et al., 2023). Here again, LLMs may offer a seemingly reasonable alternative, especially given the fact that engineering prompts is perceived as easier and cheaper than recruiting people for user research (Hämäläinen et al., 2023).

Whether generating survey or interview responses, researchers argue that LLM-generated data could be useful as it is not only cheaper and quicker to produce, but it can also potentially address sample diversity challenges (Aher et al., 2023; Argyle et al., 2023; Bail, 2024). In a recent scoping review of the efforts to use LLM-generated data for various types of social research, Agnew et al. (2024) caution that the substitution of human subjects with "homo silicus" comes into conflict with core research ethics values of representation and inclusion. They argue that study participants have important discretionary powers when participating in research, such as opting out, resisting or being able to point out misconceptions on the part of researchers. The use of LLM-generated data instead of human subjects then would shift these powers, making the resulting research inherently exclusionary. This would exacerbate already existing issues in user research, as scholars repeatedly point out LLMs tend to produce exaggerations of "stereotypical response patterns" (Simmons & Savinov, 2024; Bisbee et al., 2023) and reflect some opinions over others (Santurkar et al., 2023).

4 The Challenge of Representativeness, Privacy, Bias and Hallucinations

As scholarly excitement grows around the capacity to produce increasingly varied types of LLM-generated data, we return to the typical challenges such data are expected to address: data scarcity, privacy concerns and regulations, and lack of diversity and data bias. The papers we reviewed differed substantially in how they discussed these concerns.

The vast majority of the papers we reviewed were clearly motivated by the problem of scarcity. Results are often praised in light of the "low cost and high speed" of LLM data generation (Hämäläinen et al., 2023; Törnberg et al., 2023; Argyle et al., 2023). Agnew et al. (2024) also identify scarcity as the most common reason. As is often the case with social research, scarcity in this context is due to cost. For the most part, people are not exactly scarce — not in the way that medical images of patients affected by a rare disease can be — but they can be expensive or complicated to engage. As a result, a number of authors are enthusiastic about the possibility to scale research in social and behavioral science, where it has notoriously relied on small and unrepresentative samples (Bail, 2024; Grossmann et al., 2023; Horton, 2023). There is no doubt that LLM-produced data can surely come in any volume necessary and at a very low cost, yet it is not clear whether such scaling is, in fact, defensible or useful.

When considering the challenge of privacy and regulatory limitations, none of the papers attend to the issue, although human participants do require a growing level of privacy protection and this directly translates both in ethical limits as well as into augmented costs for data collection, processing, and storage. This is not unexpected. Existing research shows that the actual risk that LLM-based synthetic data poses via the generation of non-maliciously prompted data seems quite low (Yan et al., 2024) and it is fair to assume that LLM-based synthetic data would not fall under the protection of regulations such as GDPR and would not require complex reviews from research ethics committees.

It is the issue of data bias, diversity, and representativeness that is discussed extensively across the reviewed research. After all, for LLM-generated textual data to be viable for social science research, the capacity to produce data that is representative of populations of interest is key. What seems to be the bottom line for much of the existing research is well exemplified by Argyle et al. (2023) when they argue that "algorithmic bias" in LLMs should be treated not as a macro-level property to be corrected, but as a feature that allows the model to produce outputs that reflect expected biases in the population and different subgroups. This argument builds on the idea that, since LLMs are trained on massive amounts of online data, the data will be able to capture fine details of the social system and of the several populations in it. This assumption is often paired with the assumed ability of LLMs to be conditioned, through prompting or fine-tuning, to assume specific points of view. In this way, LLMs are able to "extract" responses that faithfully represent actual subgroups or demographics from their massive amount of training data.

Yet there are many papers that document how LLMs tend to fail to generate output that is representative of various population subgroups (Bisbee et al., 2023; Simmons & Savinov, 2024; von der Heyde et al., 2023; Cao et al. 2023). These apparently contradictory results are not surprising at this early stage. Given that the sheer amount of data needed to achieve the performance reached by recent models is hardly obtainable through curated datasets, it is difficult to know exactly what the specific model ingested as data as well as what information about specific subgroups and with what level of reliability could be extracted from it. Such capacity to create what would essentially amount to data segmentation by sub-groups appears to be one of the key arguments in favor of LLM-generated data (Argyle et al., 2023; Aher et al., 2023).

"Bias as a feature to be exploited" is a cornerstone of the idea of algorithmic fidelity. Scholars argue that biased training data and its incorporation into the model is what gives the model the ability to faithfully reproduce social groups (Argyle et al., 2023). At the same time, since their large-scale commercial deployment, biases in LLMs' outputs have been at the center of public attention (Gordon, 2023) as well as academic research (Fang et al., 2024). So much so that we have witnessed several attempts by commercial companies such as Google and OpenAI to mitigate model bias in their final output, often with mixed results (Goodman & Sandoval, 2024). Even when accepting the idea of algorithmic fidelity as unproblematic, researchers' interests and platforms' commercial plans do not seem aligned and research into the level of bias that is actually present in the final outputs of commercial models shows contradictory results (Tjuatja et al., 2023).

There are two fundamental questions — still largely unanswered — that suggest a careful approach to the idea of algorithmic fidelity and its consequent concept of algorithmic or "silicone" sampling (Argyle et al., 2023). First, what is the actual amount of bias that LLMs can reproduce? Second, what is the relation between the training data and emergent behaviors displayed by the models?

4.1 The Problem of Bias and Representativeness

Questions of bias and representativeness have spurred several studies (Bisbee et al., 2023; Simmons & Savinov, 2024; von der Heyde et al., 2023). For example, Tjuatja et al. (2023) evaluated whether nine LLMs exhibit human-like response biases in survey questionnaires. Following Törnberg et al. (2023) and Aher et al. (2023), this work leverages a framework widely used in social psychology that aims to elicit bias by changing the wording of prompts. The results demonstrated that LLMs' output is not aligned with the expected human behavior such as a "significant change in the opposite direction of known human biases, and a significant change to non-bias perturbations" (Tjuatja et al., 2023, p. 2). These observations echo research by Santurkar et al. (2023) reporting substantial differences between the views reflected by several LLMs and those of many US demographic groups, noticeable even when the model was specifically prompted to represent a particular group.

In addition to showing poor bias-reproduction the work from Tjuatja et al. (2023) also showed that LLMs that used Reinforced Learning Human Feedback (RLHF) resulted in fewer changes to question modifications as a result of response biases. Reinforced Learning Human Feedback is a specific technique that allows the model to be trained on human-feedback rather than just on data alone. This is largely used to mitigate known biases and unwanted behaviors. While the application of RLHF may result in better "products" for the general user with models that are overall more harmless and helpful (Sun, 2023), it contradicts the assertion that the inherent bias of LLMs is what affords its representativeness (Argyle et al., 2023). While the adoption of vanilla models — that have not gone through the process of RLHF — showed

https://doi.org/10.6092/issn.1971-8853/19576

some benefit, the number of researchers in the social sciences who can realistically use LLMs outside of the commercial offer, is, at the moment, quite modest.

The issue of representation gets even thornier if we consider the capacity (or lack thereof) of LLMs to address cultural diversity in human populations (Cao et al., 2023). The use of these models runs the risk of "value lock-in" (Weidinger et al., 2022) as LLMs are not able to respond to subtle changes in normative positions and opinions in the population over time. Agnew et al. (2024) point out that the use of LLMs supports notions of representation in research only in a very weak sense, unresponsive to changes in opinions, views, and preferences. As a result, studies using LLM-generated data run the risk of misrepresentation of smaller, potentially more vulnerable populations is high, essentially reproducing age-old data colonialism problems of social research (Couldry & Mejias, 2019).

4.2 The Problem of Emergent Behaviors

The second question that demands careful consideration is the tendency of LLMs towards hallucination and emergent behaviors. Transformer-based models have a well-documented tendency to hallucinate, typically defined as the production of factually incorrect yet convincing information (McKenna et al., 2023). Since in the context of LLM-based data generation the goal is not to retrieve specific information from the training data, it might seem that the problem of hallucination is not relevant to the task at hand (and this might explain why it is never mentioned in the research papers we have reviewed). Nevertheless, we would argue otherwise. Recent research from McKenna et al. (2023) shows how sentence memorization and statistical patterns in the training data are major causes of hallucinations. In both cases hallucinations are not caused by emergent properties but by "overreliance" on the sentences or the statistical patterns that have been learned from the training data. This has three possibly important consequences for data generation. First, hallucinated responses would be perfectly "believable" but, since they do not refer to any factual information, they will be harder to identify. Second, the ability of LLMs models to be effectively conditioned to reasoning outside of its training data can be limited. Third, this ability might not be equal for all the possible sub-populations researchers might want to study. This expectation that LMMs should be *segmentable*, able to reproduce multiple sub-population, is a key element in the approaches that use LLMs supporting ABMs. Here (see Törnberg et al., 2023) LLMs are explicitly asked to role-play different positions on a specific issue. We call this expectation segmentability and it is worth noticing that even if the model should preserve algorithmic fidelity to the training data, this does not imply that the model would be able to be segmented and produce data representative of various population sub-groups. This needs to be evaluated on a case-by-case basis suggesting problems of replicability and legitimacy of the resulting insights.

5 The Art and Challenge of Prompt Engineering and Evaluation

Social science research relies on robust methodological descriptions for evaluating research output and, in some fields, for ensuring replicability of results. With LLM-generated textual data, the methodological descriptions typically focus on prompt engineering as many papers argue that "proper conditioning" (Argyle, 2023) is key to ensuring fidelity of this kind of research. In many cases (with the notable exception of Park et al. 2022, 2023) prompt engineering is described as a tuning process necessary to achieve the best outputs/responses from the LLM, rather than a process with possibly profound consequences for the resulting data. Horton (2023) provides a good example of how prompts are often simply "listed".

Listing the prompts used in the research process seems to speak more to the problem of transparency of research procedures and replicability than to the problem of data production. If prompt engineering is a matter of replicability of the results, this means that the selected prompt becomes the way to unlock the model's ability to generate the desired data or the desired population. With the same prompting, the same or similar data would be produced again in the future. Yet, when investigating this specific assumption, Bisbee et al. (2023) found that generated data varies significantly both for small changes in the wording of the prompts as well as for the same prompt but asked at different moments in time. Similarly Atil et al. (2024) have reported overall instability of the output even when the conditions for deterministic behaviors are met.

If repeated prompts do not assure replicability of the research, then we have to consider how that should be documented. Lack of replicability can have many causes, from the hallucinatory nature of LLMs to the commercial nature of available platforms — platforms that are constantly updated and upgraded to offer an improved commercial product that does not need to be backward consistent. This has implications for how such data may need to be interpreted and what kinds of insights might be warranted. After all, differences due to emergent behavior have different implications to differences due to changes implemented at the platform level for commercial reasons.

Prompts lead us to consider evaluation and benchmarking. The papers we reviewed above offer different approaches to evaluating the resulting datasets for usefulness, fidelity, "faithfulness", or "believability". Where evaluations would typically cleave close to the purpose of data, they also need to be systematic and replicable (thus becoming consistent and robust research instruments). Current implementations run the gamut (thus Törnberg's essay [2024] in this special issue) but they seem to often get reduced to measures of believability — does this output *look* like human output — which does not address the issue of usefulness given the assumptions of surveys or interviews and given the fact that people are just notoriously bad at distinguishing human and AI output even for older version of LLMs (Köbis & Mossink, 2021). Nor do they consider how normative ideas of what counts as "human" may be embedded in and reproduced through use of such evaluative measures (Rhee, 2018). Where believability is useful for creating non-player characters in computer games, because their goal is only to be "believable" in interactions with players, it isn't a great measure of utility for making inferences about human responses. Just because the produced content is "believable" does not mean it has epistemic legitimacy.

This acknowledgement of different standards of "believability" draws attention to the importance the context of use and epistemological standpoint. For some social scientists, the idea that LLMs could produce "believable" material seems quite alien to the experience of, for example, conducting fieldwork to acquire relatively small amounts of qualitative data about lived experiences. For others, the material produced by an LLM may be sufficiently "believable" to be useful in a simulation. Underlying these differences are epistemological assumptions about how knowledge can and should be produced in order to say something useful about the world. Using "believability" as a way of assessing the material also tends to obscure darker questions: assessing "believability" requires a baseline in which humanness is quantified and measured, and can be used to compare outputs from LLMs. This quantification process reproduces highly problematic norms about who counts as human and why (D'Ignazio & Klein, 2023; & Gebru, 2018; Rhee, 2018).

The measure of faithfulness then may seem a better form of evaluation for some branches of the social sciences at least. Here we see replications of prior psychological or economics experiments with the idea that if LLM output aligns with what we know about how people respond to the known situation, then output produced in response to novel situations will be similarly aligned. Once again we run into several problems. The fields of psychology and economics have been going through a crisis of replicability — where scholars seem to be unable to replicate old and established experimental evidence with new studies with people. What does it say then if LLMs replicate the canon, even as it is challenged by studies with people? Many psychology and economics studies, as well as large-scale surveys in political science and demography, have been criticized for studying an unattainable ideal of the average person, statistically derived but non-existent in practice. Based on the results obtained by Bisbee et al. (2023) and by von der Heyde et al., (2023), LLMs might be in a similar situation, they can produce a set of averages — unattainable in practice.

As Horton (2023) notes, sure all models are wrong, LLMs included, but that does not mean we can't use them for thinking about what questions to ask and how to ask them. Yet it is worth reflecting on what questions might emerge given the particular notions of the average person embedded in LLMs and what kinds of questions might be left out. After all, if LLMs were to somehow produce data that might lead to fundamentally new questions, would it not by definition fail the test of faithfulness?

6 The Question of Legitimacy and Situatedness of Knowledge

While much current work acknowledges the limitations of the LLM-generated data and explores using currently available technology, there is also substantial agreement that we can expect the quality of LLM-generated textual data to improve as more complex models come online, model architectures evolve and data curation methods become ever more sophisticated (Hämäläinen et al., 2023; Törnberg et al., 2023; Park et al., 2023). This, however, does not address the epistemological issues that we have considered in this paper. Stepping back from specific technical challenges, there remain broader epistemological questions which are provoked by the increasing interest in LLM-generated data across the social sciences. Here we focus on two areas of concern: legitimacy and situated knowledge.

First, several of the papers we have analyzed build their justification on a specific version of the data scarcity argument. Data is scarce for many reasons due to the costs and challenges associated with recruiting people for research studies. LLMs promise an infinite number of quasi-human participants, lowering that cost by making quasi-human data abundant. While the cost of data is indeed a serious barrier for many researchers, it is not clear to what extent this can justify the use of LLMs without a thoughtful assessment of its epistemic legitimacy. Data scarcity is also just as likely to indicate that the research participants are unwilling to take part or have never been considered as "valid" participants before. In such cases, there may be no existing data with which to compare, or the data may be considered highly problematic. In such cases, LLM-generated data may exacerbate existing data inequalities. Following Agnew et al. (2024) critique, using such data to solve a data scarcity problem risks misrepresentation or ignores the "real" question of *why* there is no data. The research community was tempted, in the not-too-distant past, to assume that large quantities of digital data could be used as good proxies for complex social phenomena, only to find out that this was not the case (Jungherr et al., 2012).

Second, qualitative scholars have long struggled to claim the validity of their insights, especially in fields dominated by alternative epistemological positions where quantitatively produced knowledge with marginally defensible claims to generalizability was seen as the only legitimate sort. There is much excitement about the capacity to scale prior experiments and studies on limited samples through the use of LLM-generated data (Agnew et al., 2024; Bail, 2024). This stems from the underlying assumption that such data can be seen as more representative of population groups that the researchers wish to study, also known as fixing the diversity problem. Insights derived from such simulations then would only need limited substantiation in "the real world" as it were. There is nothing inherently wrong with simulations, but the challenge here is in understanding how much, and in what ways will LLM-generated outcomes differ from human answers, and in identifying how and in what ways these models may be wrong, especially if we have limited prior data with which to compare against. The danger, as we see it, is in the well-documented tendency of LLMs to produce output that reduces already expected diversity when compared to prior studies (von der Heyde et al., 2023), essentially replicating the status quo, because this is strongly embedded in the training data.

LLM transformer architectures innovate beyond the problem of the more traditional machine learning models, which in their reliance on past data for making predictions are by definition "always fighting the last war again" (Groves, 2015). Yet they too are constrained by whatever reality the training data represent — the models, after all, can only inhabit a reality described in that training data and no other. Despite the vast amount of data used for training OpenAI's GPT models, the gargantuan effort to clean training data and make them less toxic (Perrigo, 2023) speak to the desperate lack of quality or representativeness of these data. Yes, these data are the largest and "the best we got" but there is a good reason why unvarnished and uncorrected models are not made available — the mirror they hold up to humanity is profoundly terrible (Finkelstein, 2008). The models that are made available for consumption are adjusted, cleaned up, made palatable and "value-aligned" resulting in output that might create an imaginary generic average person, but who that person is, is difficult to assess. While simulating human social systems with LLMs provides intriguing insights into the models themselves, what such output might reveal about us more generally, is a question that requires cautious consideration.

How might we "situate" the knowledge produced using LLM-generated data then? The papers we cite default to lists of prompts and details of the technical setup, but arguably also require deeper considerations of where the models are wrong and what is missing, tempering the excitement with the possibility of sweeping statements about "human" behavior. What counts as "data" influences how "data" are understood, collected and processed, and are intimately connected with establishing and validating the boundaries of "proper" knowledge production (Haraway, 1988; Kitchin & Lauriault, 2018). LLM-generated data require deep considerations of fidelity — both intersectional and otherwise (Johnson & Hajisharif, 2024) — as well as of positionality, paying attention to what kinds of knowledge are made possible with the use of these data and which are foreclosed.

7 Data Must Be Cooked with Care

The idea of a computational assistant that could precisely and flexibly analyze vast amounts of complex data with quasi-qualitative skills is undoubtedly tempting for researchers, who have often struggled to adapt their methods to the growing amount and complexity of the data at their disposal. LLMs show such remarkable analytical skills as a result of the unprecedentedly large amount of data used in their training phase, but it is a leap to simply assume that these training data represent a viable proxy for the social reality behind the model. While still in its infancy, compared to other applications of LLMs, the idea of LLMs as data generators is intriguing, given the range of scholarly struggles with the complexities of data access. Some types of data may be more abundant, but they may not be easier to obtain or use.

Data is a complex and contested term, yet it has come to define the digital world we inhabit. Early debates around big data contested notions of raw data (Helmond, 2014), pointing out that data are never raw, out there, merely waiting to be collected (Bowker, 2008). Rather, data are always made, created, cooked as it were and if we are to acknowledge this, data ought to be cooked with care (Bowker, 2013). More importantly, when it comes to scientific practice, different epistemologies and methodologies "cook" data differently — seeing some methods of data generation as more legitimate than others. In many ways, synthetic data generation offers a way to create tidy, well-appointed datasets that are ultimately made specifically for this or that purpose, without the problems of cleaning messy data. Such control is one of the attractive qualities of synthetic data for many (Savage, 2023). What sort of cooking happens when generating data using LLMs? This is a much more difficult question to answer given myriad assumptions about training data, prompt processing, and emergent behaviors that must be made. Arguably, LLMs offer ease of generating data, but they provide far less control over the recipe, compared to any other approach to generating similar data from human participants.

In this essay we have looked at different examples of using LLMs for data production within the social sciences. We have discussed how LLM-generated data are similar to what we generally define as synthetic data but also where it differs. LLM-generated data aim to solve the problems of scarcity and privacy. LLMs' ability to produce large amounts of seemingly realistic data, as well as their ability to role-play various demographics with a very tenuous identifiability with the underlying training data perfectly address these needs. When it comes to bias, studies proposing the use of LLM-generated data take a different approach compared to other types of synthetic data. Rather than seeing bias as a problem that should be measured, quantified, and potentially addressed, LLM-based approaches attempt to embrace it, leveraging it, either implicitly or explicitly, as algorithmic fidelity. This difference has interesting consequences and originates from the difference in goals. While most of the recent interest towards synthetic data is driven by the need to feed more and more high-quality data into AI models to improve their performance, LLM-generated data is the output of an AI model that could potentially be used directly for research or prototyping activities. This has potential but we argue that current approaches overlook a number of important issues.

8 The Future of Data?

As we have highlighted in the sections above, the idea of using LLM-generated data for research in the social sciences is relatively new and its robustness is still disputed. The inherent complexity of LLMs, as well as their fast-paced evolution suggest caution when researchers make assumptions about the models' algorithmic fidelity or about their actual ability to be conditioned to represent various segments of the population. These are, after all, products not developed for research. Regardless of their commercial nature, which introduces additional complexity due to misaligned goals between tech companies and academic researchers, LLMs have not been developed as proxies of society. They have not been fed growing amounts of data aiming at improving their ability to represent a societal digital twin. Au contraire, many of the current trends (Saracco, 2023) that we see in the actual development of LLMs seem to suggest that the future will not be more data and more algorithmic fidelity but smaller models, trained with smaller amounts of data and fewer parameters that will still be able to score similar results in reasoning tasks.

While today LLMs can surely prove useful to produce data for prototyping research or testing initial hypotheses, their reliability should constantly be questioned and confirmed. Of course, the results we have discussed here leave open the possibility that future LLMs could be specifically designed, trained and developed for research applications. This possibility has recently been proposed by Bail (2024). Imagining large-scale LLMs developed and dedicated to research is not simple and requires a substantial change in the way social scientists approach their research tools, but it could also open up unprecedented opportunities. As many authors have noticed, what is potentially revolutionary are LLM-based models trained on large amounts of data, rather than the specific commercial implementation. Commercial solutions, while currently more advanced than open-source alternatives, come with many of the problems and limitations that we have discussed above (from unknown safeguards to fine-tuning and opaque training data). Open-source LLMs are better for ethical reasons (Spirling, 2023) and they may, at least in theory, offer better transparency, and improved control and could be based on ad-hoc training data. Yet if research-oriented open-source LLMs might be the future of LLMs for social research, it is probably a good idea to reiterate some of the key challenges they will have to face: representativeness, segmentability and data curation.

Representativeness: As many authors who have proposed the use of LLMs for data generation argue, bias that derives from biased data should not be considered a problem but as a feature of the system. Given biased training data, from a research point of view, one might want that bias to be transferred to the model's outputs. The challenge is that this is not what has been observed. While LLMs seem to be able to faithfully reproduce biases and leaning at the level of large groups they systematically fail at representing smaller groups and minorities. Algorithmic fidelity, in other words, is not stable when the system is prompted to represent certain parts of the overall population. What is more, while some types of differences may be reproduced, there is always the danger of what Johnson & Hajisharif (2024) term "intersectional hallucination" where the inherent LLM bias and built-in attempts to mitigate it might result in strange demographic configurations.

Segmentability: Directly building on the assumption of bias as a way to faithfully represent the underlying data, there is the idea that LLMs can be segmented and conditioned to represent specific sub-populations. The extent to which this is true is still unclear. While prompting and fine tuning have shown some ability to condition the results, limits have also been observed as well as a considerable amount of inconsistency even with stable prompting.

Data curation: While LLMs require, by definition, a large amount of training data, complete lack of control over what constitutes training data is problematic both for ethical and legal reasons (Rahman & Santacana, 2023). With the progress shown by smaller models (Saracco, 2023), research LLMs should carefully consider to what extent curated training data is a possibility and what would be the consequences. Over the years researchers working with *hard-to-get* data have developed considerable experience with projects of data donation (Araujo et al., 2022). This experience could be leveraged to coordinate massive collaborative efforts that would select training data not because it is available or accessible but because it has been deemed relevant. The limits and consequences of such an approach are, clearly, unknown but the ethical and legal risks of the alternatives might end up being too large for non-profit research institutions.

As things are right now these problems have been scarcely investigated and solid ways to measure them and their impact on the LLMs' ability to work as data-generation tools for social scientists have not been proposed. This should probably be a key part of any research agenda that leads to the actual development and deployment of LLMs as data generators for social sciences.

References

Abowd, J.M., & Vilhuber, L. (2008). How Protective Are Synthetic Data? In J. Domingo-Ferrer & Y. Saygın (Eds.), *Privacy in Statistical Databases* (pp. 239–246). New York, NY: Springer. https://doi.org/10.1007/978-3-540-87471-3_20

Agnew, W., Bergman, A.S., Chien, J., Díaz, M., El-Sayed, S., Pittman, J., Mohamed, S., & McKee, K.R. (2024). The Illusion of Artificial Inclusion. *arXiv*, 2401.08572. https://doi.org/10.48550/arXiv.2401.08572

Aher, G.V., Arriaga, R.I., & Kalai, A.T. (2023). Using Large Language Models to Simulate Multiple Humans and Replicate Human Subject Studies. In A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, & J. Scarlett (Eds.), *Proceedings of the 40th International Conference on Machine Learning* (pp. 337–371). https://proceedings.mlr.press/v202/aher23a.html

Almeida, G.F., Nunes, J.L., Engelmann, N., Wiegmann, A., & de Araújo, M. (2024). Exploring the Psychology of LLMs' Moral and Legal Reasoning. *Artificial Intelligence*, *333*, 104145. https://doi.org/10.1016/j.artint.2024.104145

Araujo, T., Ausloos, J., van Atteveldt, W., Loecherbach, F., Moeller, J., Ohme, J., Trilling, D., van de Velde, B., de Vreese, C., & Welbers, K. (2022). OSD2F: An Open-source Data Donation Framework. *Computational Communication Research*, *4*(2), 372–387. https://doi.org/10.5117/CCR2022.2.001.ARAU

Argyle, L.P., Busby, E.C., Fulda, N., Gubler, J.R., Rytting, C., & Wingate, D. (2023). Out of One, Many: Using Language Models to Simulate Human Samples. *Political Analysis*, *31*(3), 337–351. https://doi.org/10.1017/pan.2023.2

Atil, B., Chittams, A., Fu, L., Ture, F., Xu, L., & Baldwin, B. (2024). LLM Stability: A Detailed Analysis with Some Surprises. *arXiv*, 2408.04667. https://doi.org/10.48550/arXiv.2408.04667

Bail, C.A. (2024). Can Generative AI Improve Social Science?. *Proceedings of the National Academy of Sciences*, *PNAS*, *121*(21), e2314021121. https://doi.org/10.1073/pnas.2314021121

Belgodere, B., Dognin, P., Ivankay, A., Melnyk, I., Mroueh, Y., Mojsilovic, A., Navratil, J., Nitsure, A., Padhi, I., Rigotti, M., Ross, J., Schiff, Y., Vedpathak, R., & Young, R.A. (2023). Auditing and Generating Synthetic Data with Controllable Trust Trade-offs. *arXiv*, 2304.10819. https://doi.org/10.48550/arXiv.2304.10819

Bisbee, J., Clinton, J., Dorff, C., Kenkel, B., & Larson, J. (2023). Synthetic Replacements for Human Survey Data? The Perils of Large Language Models. *SocArXiv*, May 4. https://doi.org/10.31235/osf.io/5ecfa

Brand, J., Israeli, A., & Ngwe, D. (2023). Using LLMs for Market Research (Harvard Business School Marketing Unit Working Paper No. 23-062). *Social Science Research Network*. https://doi.org/10.2139/ssrn.4395751

Breum, S.M., Egdal, D.V., Mortensen, V.G., Møller, A.G., & Aiello, L.M. (2023). The Persuasive Power of Large Language Models. *arXiv*, 2312.15523. https://doi.org/10.48550/arXiv.2312.15523

Box, G.E.P. (1976). Science and Statistics. *Journal of the American Statistical Association*, *71*(356), 791–799. https://doi.org/10.1080/01621459.1976.10480949

Bowker, G.C. (2008). *Memory Practices in the Sciences*. Cambridge, MA: MIT Press.

Bowker, G.C. (2013). Data Flakes: An Afterword to "Raw Data" Is an Oxymoron. In L. Gitelman (Ed.), *"Raw Data" Is an Oxymoron* (pp. 167–172). Cambridge, MA: MIT Press.

Buolamwini, J., & Gebru, T. (2018). Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. In S.A. Friedler & C. Wilson (Eds.), *Proceedings of the 1st Conference on Fairness, Accountability, and Transparency*, *PMLR*, *81*, 77–91. http://proceedings.mlr.press/v81/buolamwini18a/buolamwini18a.pdf

Cao, Y., Zhou, L., Lee, S., Cabello, L., Chen, M., & Hershcovich, D. (2023). Assessing Cross-Cultural Alignment between ChatGPT and Human Societies: An Empirical Study. In S. Dev, V. Prabhakaran, D. Adelani, D. Hovy, & L. Benotti (Eds.), *Proceedings of the First Workshop on Cross-Cultural Considerations in NLP* (pp. 53–67). Singapore: Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.c3nlp-1.7)

Couldry, N., & Mejias, U.A. (2019). Data Colonialism: Rethinking Big Data's Relation to the Contemporary Subject. *Television & New Media*, *20*(4), 336–349. https://doi.org/10.1177/1527476418796632

Cui, R., Lee, S., Hershcovich, D., & Søgaard, A. (2023). What Does the Failure to Reason with "Respectively" in Zero/Few-Shot Settings Tell Us about Language Models?. In A. Rogers, J. Boyd-Graber, & N. Okazaki (Eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics, Vol.1* (pp. 8786–8800). Singapore: Association for Computational Linguistics.

Demuro, E., & Gurney, L. (2024). Artificial Intelligence and the Ethnographic Encounter: Transhuman Language Ontologies, or What It Means "To Write like a Human, Think like a Machine". *Language & Communication*, *96*, 1–12. https://doi.org/10.1016/j.langcom.2024.02.002

D'Ignazio, C., & Klein, L.F. (2023). *Data Feminism*. Cambridge, MA: MIT Press.

Dillion, D., Tandon, N., Gu, Y., & Gray, K. (2023). Can AI Language Models Replace Human Participants? *Trends in Cognitive Sciences*, *27*(7), 597–600. https://doi.org/10.1016/j.tics.2023.04.008

Eigenschink, P., Reutterer, T., Vamosi, S., Vamosi, R., Sun, C., & Kalcher, K. (2023). Deep Generative Models for Synthetic Data: A Survey. *IEEE Access*, *11*, 47304–47320. https://doi.org/10.1109/ACCESS.2023.3275134

Elsayed-Ali, S., Bonsignore, E., & Chan, J. (2023). Exploring Challenges to Inclusion in Participatory Design From the Perspectives of Global North Practitioners. In J. Nichols (Ed.), *Proceedings of the ACM on Human-Computer Interaction* (p. 7). New York, NY: Association for Computing Machinery https://doi.org/10.1145/3579606

Fang, X., Che, S., Mao, M., Zhang, H., Zhao, M., & Zhao, X. (2024). Bias of AI-generated Content: An Examination of News Produced by Large Language Models. *Scientific Reports*, *14*(1), 5224, 1–20. https://doi.org/10.1038/s41598-024-55686-2

Figueira, A., & Vaz, B. (2022). Survey on Synthetic Data Generation, Evaluation Methods and GANs. *Mathematics*, *10*(15), 1–41. https://doi.org/10.3390/math10152733

Finkelstein, S. (2008). Google, Links, and Popularity versus Authority. In J. Turow & L. Tsui (Eds.), *The Hyperlinked Society: Questioning Connections in the Digital Age* (pp. 104–120). Ann Arbor, MI: University of Michigan Press.

Goodman, J.D., & Sandoval, E. (2024). Google Chatbot's A.I. Images Put People of Color in Nazi-era Uniforms. *The New York Times*, 22 February. https://www.nytimes.com/2024/02/22/technology/google-gemini-german-uniforms.html

Gordon, R. (2023). Large Language Models Are Biased. Can Logic Help Save Them? *MIT News*, 3 March. https://news.mit.edu/2023/large-language-models-are-biased-can-logic-help-save-them-0303

Greenwood, J. (2018). How Would People Behave in Milgram's Experiment Today. *Behavioral Scientist*, 24 July. https://behavioralscientist.org/how-would-people-behave-in-milgrams-experiment-today

Grossmann, I., Feinberg, M., Parker, D.C., Christakis, N.A., Tetlock, P.E., & Cunningham, W.A. (2023). AI and the Transformation of Social Science Research. *Science*, *380*(6650), 1108–1109. https://doi.org/10.1126/science.adi1778

Groves, C. (2015). Logic of Choice or Logic of Care? Uncertainty, Technological Mediation and Responsible Innovation. *NanoEthics*, *9*(3), 321–333. https://doi.org/10.1007/s11569-015-0238-x

Hämäläinen, P., Tavast, M., & Kunnari, A. (2023). Evaluating Large Language Models in Generating Synthetic HCI Research Data: A Case Study. In A. Schmidt, K. Väänänen, T. Goyal, P.O. Kristensson, A. Peters, S. Mueller, J.R. Williamson, & M.L. Wilson (Eds.), *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (pp. 1–19). New York, NY: Association for Computing Machinery. https://doi.org/10.1145/3544548.3580688

Haraway, D. (1988). Situated Knowledges: The Science Question in Feminism and the Privilege of Partial Perspective. *Feminist Studies*, *14*(3), 575–599. https://doi.org/10.2307/3178066

Helmond, A. (2014). "Raw Data" Is an Oxymoron. *Information, Communication & Society*, *17*(9), 1171–1173. https://doi.org/10.1080/1369118X.2014.920042

Henrich, J., Heine, S.J., & Norenzayan, A. (2010). The Weirdest People in the World? *Behavioral and Brain Sciences*, *33*(2–3), 61–83. https://doi.org/10.1017/S0140525X0999152X

Horton, J.J. (2023). Large Language Models as Simulated Economic Agents: What Can We Learn from Homo Silicus? (Working Paper 31122). *National Bureau of Economic Research*. https://doi.org/10.3386/w31122

Jacobsen, B.N. (2023). Machine Learning and the Politics of Synthetic Data. *Big Data & Society*, *10*(1), 20539517221145372, 1–12. https://doi.org/10.1177/20539517221145372

Jakesch, M., Hancock, J.T., & Naaman, M. (2023). Human Heuristics for AI-Generated Language Are Flawed. *Proceedings of the National Academy of Sciences*, *PNAS*, *120*(11), e2208839120, 1–7. https://doi.org/10.1073/pnas.2208839120

Jansen, B.J., Jung, S., & Salminen, J. (2023). Employing Large Language Models in Survey Research. *Natural Language Processing Journal*, *4*, 100020, 1–7. https://doi.org/10.1016/j.nlp.2023.100020

Johnson, E., & Hajisharif, S. (2024). The Intersectional Hallucinations of Synthetic Data. *AI & Society*, 1–3. https://doi.org/10.1007/s00146-024-02017-8

Jones, C.R., Trott, S., & Bergen, B. (2023). Epitome: Experimental Protocol Inventory for Theory of Mind Evaluation. *Proceedings of the First Workshop on Theory of Mind in Communicating Agents*, *PMLR*, *202*. https://openreview.net/pdf?id=e5Yky8Fnvj

Jordon, J., Szpruch, L., Houssiau, F., Bottarelli, M., Cherubin, G., Maple, C., Cohen, S.N., & Weller, A. (2022). Synthetic Data. What, Why and How? *arXiv*, 2205.03257. https://doi.org/10.48550/arXiv.2205.03257

Jungherr, A., Jürgens, P., & Schoen, H. (2012). Why the Pirate Party Won the German Election of 2009 or the Trouble with Predictions: A Response to Tumasjan, A., Sprenger, T.O., Sander, P.G., & Welpe, I.M. "Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment". *Social Science Computer Review*, *30*(2), 229–234. https://doi.org/10.1177/0894439311404119

Kitchin, R., & Lauriault, T.P. (2018). Toward Critical Data Studies: Charting and Unpacking Data Assemblages and Their Work. In J. Thatcher, J. Eckert & A. Shears (Eds.), *Thinking Big Data in Geography: New Regimes, New Research* (pp. 3–20). Lincoln, NE: University of Nebraska Press.

Köbis, N., & Mossink, L.D. (2021). Artificial Intelligence versus Maya Angelou: Experimental Evidence that People Cannot Differentiate AI-generated from Human-written Poetry. *Computers in Human Behavior*, *114*, 106553, 1–13. https://doi.org/10.1016/j.chb.2020.106553

McKenna, N., Li, T., Cheng, L., Hosseini, M., Johnson, M., & Steedman, M. (2023). Sources of Hallucination by Large Language Models on Inference Tasks. *Findings of the Association for Computational Linguistics: EMNLP 2023* (pp. 2758–2774). Singapore: Association for Computational Linguistics https://doi.org/10.18653/v1/2023.findings-emnlp.182

Moats, D., & Seaver, N. (2019). "You Social Scientists Love Mind Games": Experimenting in the "Divide" between Data Science and Critical Algorithm Studies. *Big Data & Society*, *6*(1), 2053951719833404, 1–11. https://doi.org/10.1177/2053951719833404

Møller, A.G., Pera, A., Dalsgaard, J., & Aiello, L. (2024). The Parrot Dilemma: Human-labeled vs. LLM-augmented Data in Classification Tasks. In Y. Graham, & M. Purver (Eds.), *Pro-*

ceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 2: Short Papers) (pp. 179–192). Singapore: Association for Computational Linguistics. https://aclanthology.org/2024.eacl-short.17/

Motoki, F., Pinho Neto, V., & Rodrigues, V. (2024). More Human than Human: Measuring ChatGPT Political Bias. *Public Choice*, *198*(1–2), 3–23. https://doi.org/10.1007/s11127-023-01097-2

Nikolenko, S.I. (2021). *Synthetic Data for Deep Learning*. Cham: Springer International Publishing.

Park, J.S., Popowski, L., Cai, C., Morris, M.R., Liang, P., & Bernstein, M.S. (2022). Social Simulacra: Creating Populated Prototypes for Social Computing Systems. In M. Agrawala, J.O. Wobbrock, E. Adar, & V. Setlur (Eds.), *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology* (pp. 1–18). New York, NY: Association for Computing Machinery. https://doi.org/10.1145/3526113.3545616

Park, J.S., O'Brien, J., Cai, C.J., Morris, M.R., Liang, P., & Bernstein, M.S. (2023). Generative Agents: Interactive Simulacra of Human Behavior. In S. Follmer, J. Han, J. Steimle, & N. Henry Riche (Eds.), *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology* (pp. 1–22). New York, NY: Association for Computing Machinery. https://doi.org/10.1145/3586183.3606763

Perrigo, B. (2023). OpenAI Used Kenyan Workers on Less Than $2 Per Hour: Exclusive. *Time*, 18 January. https://time.com/6247678/openai-chatgpt-kenya-workers/

Phelps, S., & Russell, Y.I. (2023). Investigating Emergent Goal-Like Behaviour in Large Language Models Using Experimental Economics. *arXiv*, 2305.07970. https://doi.org/10.48550/arXiv.2305.07970

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language Models Are Unsupervised Multitask Learners. *OpenAI*, *1*(8), 1–24. https://api.semanticscholar.org/CorpusID:160025533

Raghunathan, T.E. (2021). Synthetic Data. *Annual Review of Statistics and Its Application*, *8*(1), 129–140. https://doi.org/10.1146/annurev-statistics-040720-031848

Rahman, N., & Santacana, E. (2023). Beyond Fair Use: Legal Risk Evaluation for Training LLMs on Copyrighted Text. *Proceedings of the 40th International Conference on Machine Learning* (pp. 1–5). https://blog.genlaw.org/CameraReady/57.pdf

Rhee, J., (2018). *The Robotic Imaginary: The Human and the Price of Dehumanized Labor*. Minneapolis, MN: University of Minnesota Press.

Rossetti, G., Stella, M., Cazabet, R., Abramski, K., Cau, E., Citraro, S., Failla, A., Improta, R., Morini, V., & Pansanella, V. (2024). Y Social: An LLM-powered Social Media Digital Twin. *arXiv*, 2408.00818. https://doi.org/10.48550/arXiv.2408.00818

Salminen, J., Guan, K.W., Jung, S.G., & Jansen, B. (2022). Use Cases for Design Personas: A Systematic Review and New Frontiers. In S. Barbosa, C. Lampe, C. Appert, D.A. Shamma, S. Drucker, J. Williamson, & K. Yatani (Eds.), *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (pp. 1–21). New York, NY: Association for Computing Machinery. https://doi.org/10.1145/3491102.3517589

Samuelson, W., & Zeckhauser, R. (1988). Status Quo Bias in Decision Making. *Journal of Risk and Uncertainty*, *1*, 7–59. https://doi.org/10.1007/BF00055564

Santurkar, S., Durmus, E., Ladhak, F., Lee, C., Liang, P., & Hashimoto, T. (2023). Whose Opinions Do Language Models Reflect? In A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, & J. Scarlett (Eds.), *Proceedings of the 40th International Conference on Machine Learning, PMLR*, *202*, 29971–30004. https://proceedings.mlr.press/v202/santurkar23a.html

Saracco, R. (2023). How Much Bigger Can/Should LLMs Become?. *IEEE Future Directions*, April 24. Retrieved from https://cmte.ieee.org/futuredirections/2023/04/24/how-much-bigger-can-should-llms-become/

Savage, N. (2023). Synthetic Data Could Be Better than Real Data. *Nature*, d41586-023-01445-01448, 27 April. https://www.nature.com/articles/d41586-023-01445-8

Schaeffer, R., Miranda, B., & Koyejo, S. (2024). Are Emergent Abilities of Large Language Models a Mirage? *Proceedings of the 37ìth International Conference on Neural Information Processing Systems* (pp. 55565–55581). https://doi.org/10.48550/arXiv.2304.15004

Simmons, G., & Savinov, V. (2024). Assessing Generalization for Subpopulation Representative Modeling via In-Context Learning. *arXiv*, 2402.07368. https://doi.org/10.48550/ARXIV.2402.07368

Sin, J., Franz, R.L., Munteanu, C., & Barbosa Neves, B. (2021). Digital Design Marginalization: New Perspectives on Designing Inclusive Interfaces. In Y. Kitamura & A. Quigley (Eds.), *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (pp. 1–11). https://doi.org/10.1145/3411764.3445180

Spirling, A. (2023). Why Open-source Generative AI Models Are an Ethical Way Forward for Science. *Nature*, *616*((7957), 413–413. https://doi.org/10.1038/d41586-023-01295-4

Sun, H. (2023). Reinforcement Learning in the Era of LLMs: What Is Essential? What Is Needed? An RL Perspective on RLHF, Prompting, and Beyond. *arXiv*, 2310.06147. https://doi.org/10.48550/arXiv.2310.06147

Tjuatja, L., Chen, V., Wu, S.T., Talwalkar, A., & Neubig, G. (2023). Do LLMs Exhibit Human-like Response Biases? A Case Study in Survey Design. *arXiv*, 2311.04076. https://doi.org/10.48550/ARXIV.2311.04076

Törnberg, P., Valeeva, D., Uitermark, J., & Bail, C. (2023). Simulating Social Media Using Large Language Models to Evaluate Alternative News Feed Algorithms. *arXiv*, 2310.05984. https://doi.org/10.48550/arXiv.2310.05984

Törnberg, P. (2024). Best Practices for Text Annotation with Large Language Models. *Sociologica*, *18*(2), 67–85. https://doi.org/10.6092/issn.1971-8853/19461

van der Schaar, M., & Qian, Z. (2023). *AAAI Lab for Innovative Uses of Synthetic Data*. Association for the Advancement of Artificial Intelligence. https://www.vanderschaar-lab.com/wp-content/uploads/2022/08/AAAI_Synthetic-Data-Tutorial.pdf von der Heyde, L., Haensch, A.-C., & Wenz, A. (2023). Assessing Bias in LLM-Generated Synthetic Datasets: The Case of German Voter Behavior. *SocArXiv*. https://EconPapers.repec.org/RePEc:osf:socarx:97r8s

Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., Yogatama, D., Bosma, M., Zhou, D., Metzler, D., Chi, E.H., Hashimoto, T., Vinyals, O., Liang, P., Dean, J., & Fedus, W. (2022). Emergent Abilities of Large Language Models. *arXiv*, 2206.07682. https://doi.org/10.48550/arXiv.2206.07682

Weidinger, L., Uesato, J., Rauh, M., Griffin, C., Huang, P., Mellor, J., Glaese, A., Cheng, M., Balle, B., Kasirzadeh, A., Biles, C., Brown, S., Kenton, Z., Hawkins, W., Stepleton, T., Birhane, A., Hendricks, L.A., Rimell, L., Isaac, W., … Gabriel, I. (2022). Taxonomy of Risks Posed by Language Models. *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency* (pp. 214–229). https://doi.org/10.1145/3531146.3533088

Whitney, C.D., & Norman, J. (2024). Real Risks of Fake Data: Synthetic Data, Diversity-washing and Consent Circumvention. *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency* (pp. 1733–1744). New York, NY: Association for Computing Machinery. https://doi.org/10.1145/3630106.3659002

Wu, Z., Peng, R., Han, X., Zheng, S., Zhang, Y., & Xiao, C. (2023). Smart Agent-Based Modeling: On the Use of Large Language Models in Computer Simulations (arXiv: 2311.06330). *arXiv*. https://doi.org/10.48550/arXiv.2311.06330

Yan, B., Li, K., Xu, M., Dong, Y., Zhang, Y., Ren, Z., & Cheng, X. (2024). On Protecting the Data Privacy of Large Language Models (LLMs): A Survey. *arXiv*, 2403.05156. https://doi.org/10.48550/ARXIV.2403.05156

Luca Rossi – NERDS research group, Department of Digital Design, IT University of Copenhagen (Denmark)
https://orcid.org/0000-0002-3629-2039 | lucr@itu.dk
https://pure.itu.dk/da/persons/luca-rossi
Luca Rossi is an Associate Professor of Digital Media and Networks at the Department of Digital Design of IT, University of Copenhagen (Denmark). He coordinates the Human Centered Data Science research group, and he is member of the Networks Data and Society (NERDS) research group. He teaches Network analysis and Digital Data Analysis.

Katherine Harrison – Department of Thematic Studies – Gender Studies, Linköping University (Sweden)
https://orcid.org/0000-0002-8325-4051
https://liu.se/en/employee/katha38
Katherine Harrison, Ph.D., is an Associate Professor in Gender Studies at Linköping University (Sweden). Her research sits at the intersection of Science & Technology Studies, media studies, and feminist theory, bringing critical perspectives on knowledge production to studies of different digital technologies.

Irina Shklovski – Department of Computer Science, Department of Communication, University of Copenhagen (Denmark); Department of Thematic Studies – Gender Studies, Linköping University (Sweden)
https://orcid.org/0000-0003-1874-0958
https://researchprofiles.ku.dk/en/persons/irina-shklovski
Irina Shklovski is a Professor of Communication and Computing in the Department of Computer Science and the Department of Communication at the University of Copenhagen (Denmark). She holds a WASP-HS visiting professorship at Linköping University (Sweden). Her research areas include speculative AI futures, AI ethics, data quality, synthetic data, explainability, privacy, and creepy technologies.