Reproducibility Under Threat: Proposing a Framework for Reliable LLM-Research in Psychology and Computational Social Science

Kevin D. Kiy[1], Alexander Porshnev[1], Dounia Lakhzoum[1], Dermot Lynott[1], Diarmuid O'Donoghue[2], Manokamna Singh[2]

[1]Department of Psychology, Maynooth University, Maynooth, Ireland
[2]Department of Computer Science, Maynooth University, Maynooth, Ireland
kevin.kiy.2024@mumail.ie, {alexander.porshnev, dounia.lakhzoum, dermot.lynott, diarmuid.odonoghue}@mu.ie, manokamna.singh.2023@mumail.ie

Abstract

The integration of artificial intelligence (AI), particularly large language models (LLMs), into research across the social sciences has accelerated innovation but also introduced significant challenges to reproducibility - a cornerstone of scientific integrity. In this review of scientific practices, we examine the reproducibility crisis in AI-driven research with a focus on psychology, identifying common pitfalls, reviewing proposed solutions, and advocating for best practices. Common pitfalls in current practices in the social sciences are identified and highlighted through synthesized research scenarios, such as: (1) using inaccessible datasets or language models with restricted access, (2) treating black-box API outputs as stable observations ignoring updates and hidden changes, (3) producing single runs for measurements instead of stochastic draws for aggregated performances, (4) failing to report full LLM version, prompting, and sampling parameters, and (5) opaque training and fine-tuning of LLMs. Our recommended practices include precisely documenting the model used, fixing all inference parameters, using automation and scripts to control prompts, context, and outputs, and standardizing the environment and API conditions. By embracing transparency and methodological rigour, we can transform the challenges of AI-driven research into opportunities for more robust and impactful science, ensuring that innovation never comes at the cost of credibility.

Keywords: Reproducibility, Open Science, Responsible AI Practices

1. Introduction

In recent years, the study of large language models (LLMs) in psychology and cognitive science has seen a surge in popularity (Demszky et al., 2023). These implementations range from using LLMs to simulate human participants in psychological studies (Dillon et al., 2023; Shiffrin, 2023) to serving as sandboxes for testing cognitive theories (Strachan et al., 2024; Niu et al., 2024). However, this rapid adoption raises concerns about the emergence of a new reproducibility crisis as many studies fail to account for common pitfalls inherent to LLMs: (1) treating black-box API outputs as stable observations ignoring updates and hidden changes (Morishige & Koshihara, 2025); (2) failing to report full LLM prompting and sampling parameters (Mitchell et al., 2019; Kapoor et al., 2024); (3) producing single runs for measurements instead of stochastic draws for aggregated performances (Guo et al., 2025). Considering that these issues prevent the reconstruction of the exact computational conditions under which the results were obtained, they represent a direct threat to reproducibility. In this paper, we review these pitfalls and propose concrete solutions for researchers, including reproducible methods and reporting standards.

2. Background

In the following section, we highlight current concerns surrounding reproducibility in research within the broader context of the reproducibility crisis and examine how LLM-based research introduce novel challenges that risk undermining previous efforts to promote open research practices.

2.1 Reproducibility in research

Following widespread replication failures across multiple fields, the research community has become acutely aware of issues of reproducibility in scientific research. In a survey conducted by Nature of 1,500 researchers, 90% of respondents affirmed they believe there is indeed a reproducibility crisis primarily due to Questionable Research Practices (QRPs; Baker, 2016; Bakker et al., 2021; Fanelli, 2018, John et al., 2012; Smaldino & McElreath, 2016). These questionable practices include analytical ambiguity (e.g., p-haking, HARKing or Hypothesizing After the Results are Known; cherry-picking data, choosing inadequate sample sizes); selective reporting (e.g., suppressing non-significant findings), and insufficient transparency and documentation (e.g., lack of details in reporting methods and results, preprocessing pipelines, failure to share data, materials, or code). In response, substantial efforts have been devoted to addressing these issues through the adoption of Open Research Practices (ORPs; e.g., preregistration, data and code sharing), developing detailed documentation guidelines and replication initiatives aimed at ensuring rigorous, reliable, and robust research. These efforts have been particularly prominent in the social and cognitive sciences after highly publicised replication failures sparked a reform

Proceedings of The 3rd Workshop on Natural Language Processing for Political Sciences (PoliticalNLP 2026) @ LREC 2026, pages 171–179
16 May 2026, Palma, Mallorca (Spain).

movement and positioned these fields at the forefront of developing and implementing open practices (Nelson et al., 2018; Open Science Collaboration, 2015; Munafo et al., 2020). Following a decade of self-assessment and meta-scientific investigation, these initiatives have resulted in measurable improvements in transparency, reporting quality, methodological rigour and have ushered a new open science frontier (Nosek et al., 2015; Munafo et al., 2017; Christensen et al., 2018). Despite these substantial improvements, emerging research paradigms based on large language models (LLMs) introduce new structural challenges that risk reversing these gains (Hutson, 2018; Pineau et al., 2021). LLM research represents a fast-paced, resource-intensive and transformative paradigm, holding substantial promise for opening new frontiers in the social and cognitive sciences (Wei et al., 2022). Naturally, these tools have gained considerable popularity in recent years given their capacity to simulate human-like linguistic behaviour or explore emergent cognitive phenomena. Despite these promises, certain inherent properties of LLM-based research introduce structural challenges for reproducibility. First, given the restricted access to core components including training data, optimization procedures, and weight distributions, the opportunity for independent verification and systemic replication is limited (Paullada et al., 2021; Liang, 2022). In addition, the probabilistic nature of LLM outputs, combined with undocumented third-party updates undermine the assumption of stable experimental observations across studies (Chen et al., 2023). Finally, the absence of standardized frameworks for reporting impedes exhaustive and transparent documentation of processes, pipelines and findings (Mitchell et al., 2019). Taken together, these structural constraints frame a reproducibility landscape where the safeguards developed over the past decade become under threat. Beyond methodological reproducibility, the rapid adoption of large language models in academic research raises broader concerns about responsible AI, spanning economic, ecological, and ethical dimensions (Bender et al., 2021; Bommasani et al., 2021; Strubell et al., 2019).

2.2 Responsible AI in academia

To ensure that AI-usage in academia is responsible, one must account for the economic, ecological, and ethical considerations in regard to model selection and methodology (Waelen & van Wynsberghe, 2025). The economic incentives to commercialize AI tools often conflict with open science principles, resulting in open-source models being behind proprietary AI models in dataset size (Hartmann & Henkel, 2020), computational capacity (Ali et al., 2025), and general societal influence (Pei & Huang, 2025; Bommasani et al., 2021). While more cost-effective smaller models can outperform less open large language models when trained for domain-specific tasks (Porshnev et al., 2024), researchers aiming to use the most powerful tools available to them are faced with the costs and restricted access of closed-source AI systems. Beyond the economic costs of using proprietary large language models, there are further ecological consequences to be mindful of. The International Energy Agency (2025) reports that data centers for AI-development accounted for ~1.5% of global electricity use in 2024 and will go up to an estimated use of 3% by 2030. Further, an estimated amount of CO_2 equivalent to the emissions of a trans-American flight are emitted for a singular training session of a large language model (Strubell et al., 2019). While these ecological consequences are worthy of moral concern in and of themselves, there are further ethical considerations to be taken into account for AI usage and development. Language models have been found to exhibit human-like biases that can propriate harmful stereotypes (Caliskan et al., 2017; Lynott et al. 2019; Bhatia & Walasek 2023). Additional biases beyond those picked up by language models from the linguistic data that they are trained with can be caused by human decision-making in the development and fine-tuning process of AI products (Bommasani et al., 2021), like underrepresenting a language (Kreutzer et al., 2022), or minority (Venkit et al., 2022; Hutchinson et al., 2020) in the training data. The black-box nature of most closed-source AI products makes it nearly impossible to be aware of all biases present in the model one is currently using (Hassija et al., 2024), which limits the amount of control and stability a researcher can ensure for his AI-assisted investigation, and therefore further debilitates the notion of responsibility for such research endeavours. For instance, social scientists wanting to utilize LLM-technology for their (political) discourse analysis research need to be especially cautious of using black-box models for their studies, as social biases (Rettenberger et al., 2024) or political stances (Walker & Angst, 2025) inherent to the model strongly influence the resulting outputs one receives from the machine (Feng et al., 2023). While these considerations of responsibility are broad and apply more generally to all academic disciplines, more specific pitfalls in AI-driven research can be pinpointed in the scientific outputs in psychology and the social sciences.

3. Highlighting common mistakes: Non-reproducible LLM-based research scenarios

The following research scenarios describe practices that are commonly found in AI-assisted research in psychology and the social sciences and usually hinders or fully invalidates the reproducibility of these scientific outputs.

3.1 Using inaccessible datasets or language models

Researchers often rely on proprietary datasets or closed-source language models (e.g., GPT-5, Claude) without providing access to the data or model weights. This practice renders the study irreproducible, as independent verification or replication is impossible without the original inputs or tools. Without access to the dataset or model, other researchers cannot verify the findings, test alternative hypotheses, or build on the work. This trend stands as a polar opposite to the movement of open science in social science and research as a whole, where it has become standard practice to share anonymized datasets and materials (via repositories like OSF, see Foster & Deardorff, 2017) to enable replication. Scientists should focus on the openness of their language models just as much as they do for the rest of their research. One could do so by consulting the Artificial Intelligence Openness Index (Artificial Analysis, 2026). It introduces a standard metric for assessing and comparing LLM openness, composed firstly of model availability and access to weights. Secondly, it assesses model transparency assessing training data and disclosure of the training methodology. This index identifies models scoring poorly on the openness index including (at the time of writing, ver. 1.0) GPT-5 mini, o3 and GPT-5 Nano all scoring just 6, while the most open models include Molmo 7B-D and Olmo 3 7B Think scoring 89. This highlights that even the process of selecting LLM for comparison can be a complicated process, with many of the most popular and arguably influential models showing few openness traits. Conversely, many of the most open models are not only less-well known but they include some of the smaller models (7B parameters models) which might fail to challenge the newer and more powerful models.

3.2 Treating black-box API outputs as stable observations

Outputs from LLM APIs (e.g., OpenAI's API) are often treated as static, reliable measurements. However, these outputs can vary due to unseen model updates, temperature settings, or hidden backend changes, which are rarely documented or announced. API outputs are not immutable; they can shift over time due to model retraining or infrastructure changes. Studies that do not account for this variability risk reporting findings that cannot be replicated, even if the same prompts are reused (Morishige & Koshihara, 2025). In experimental psychology, researchers control for confounds like participant fatigue or environmental noise to ensure stable measurements. To allow for more control and enable stable and therefore reproducible outputs, researchers should introduce a measure of stability when reporting their results (Atil, et al., 2024) or opt to use models where the degree of variability can be controlled, e.g. by locally hosting accessible models (using a platform like huggingface.co).

3.3 Producing single runs for measurements

Many studies report results from a single LLM output per prompt, treating it as a definitive answer. This ignores the stochastic nature of LLMs, where the same prompt can yield different responses due to sampling variability (Guo et al., 2025). Single-run outputs are analogous to drawing conclusions from a single participant in a survey. Without accounting for variability, findings may reflect noise rather than robust patterns, and replication attempts may yield conflicting results (Belz, et al., 2021). In quantitative research, statistical power is achieved through sufficient sample sizes. For LLM-research, this could mean having a sufficient number of outputs for each input and testing the stability of the results (Atil, et al., 2024).

3.4 Failing to report full LLM prompting and sampling parameters

Researchers often omit critical details about how prompts were constructed or how model parameters (e.g., temperature, max tokens, frequency penalties) were set. This omission leaves gaps in the methodology that prevent exact replication (Mitchell et al., 2019; Kapoor et al., 2024). Without full transparency, even minor differences in prompting or parameter settings can lead to divergent results. For example, a prompt phrased as "List the causes of X" may yield different outputs than "What are the causes of X?" In qualitative research, interview protocols and coding schemes are meticulously documented to ensure consistency. In the same way, LLM prompting and sampling parameters should be documented, pre-registered, and openly made available to the scientific community to allow for exact replication and understanding of each methodological step and decision (Magnusson et al. 2023).

3.5 Opaque training and fine-tuning parameters

Similarly to the issues described in 3.4, studies that fine-tune LLMs often neglect to report hyperparameters, training data composition, or evaluation metrics. This opacity makes it impossible to assess whether the model's performance is due to the method or idiosyncratic choices. Without these details, other researchers cannot replicate the fine-tuning process or validate the model's generalizability (Atil, et al., 2024). For example, a model fine-tuned on a biased dataset may appear effective in isolation but fail in real-world applications. The following chapter outlines detailed best practices for LLM-usage in computational social science to avoid the pitfalls described here and above.

4. Identifying best practices for LLM-usage in Computational Social Sciences

Best practices for reproducible usage of large language models (LLMs) in social-science research closely mirror established best practices for reproducible code, transparent data management, and well-documented analysis pipelines. These are essential points to ensure that results can be verified, replicated, and meaningfully interpreted by others. It is worth mentioning that the task of identifying best practices for the reproducibility of AI-assisted research is quite different from the topic of appropriate use of LLMs in general research practice, e.g. that LLMs could be used as well for hypothesis and text generation (Chairs, 2023).

Abdurahman and coauthors suggested that reviewers, when evaluating papers that use large language models (LLMs), should pay attention to the following practices (which are undoubtedly equally important for authors):

- Provide replication materials:
  Code, prompts, model parameters, fine-tuning data, study material (e.g., questionnaires), and human-validation data; Discuss strategies to account for LLM randomness.
- Check model stability:
  Check if model changes over time and if the model version you used is accessible over time.
- Validate:
  Check LLM outputs against human data or other ground truth; check robustness to different prompt strategies and model settings; unbiased data processing and error handling (Abdurahman et al., 2025).

One of the most visible concerns relates to the instability of cloud-hosted LLMs. Commercial providers may update models without notice, altering their weights, training data mixtures, or alignment procedures. Even small updates can lead to measurable changes in outputs, meaning that results generated at one point in time may not be replicable later. For this reason, researchers should strongly consider using open-access or locally hosted models whenever feasible. When a model can be downloaded, version-pinned, and archived ideally with a persistent identifier other scholars are better positioned to reproduce the exact computational conditions under which findings were generated. Progress with infrastructure eases the application of large language models, e.g. GPT4All (https://docs.gpt4all.io/) or Ollama (https://ollama.com/) can help researchers to run open access models locally from https://huggingface.co/. Due to the fast-paced environment of LLM-development, it is hard to recommend a particular model agnostic to the respective research design and study aim, but for general quality and robustness of analysis it is often most optimal for researchers to test small language models from different sources (e.g. Mistral 7B, LLama 3.1 7B) as they provide a good trade-off between openness, capability on (political) text analysis tasks, and computational feasibility for typical academic compute environments.

Another well-known issue is the strong dependence of outputs on prompt wording. Minor changes in phrasing, order, formatting, or instruction style can produce substantively different responses. In social-science contexts where interpretation, classification, or coding decisions may feed directly into statistical analyses such sensitivity is not trivial. Prompts should therefore be treated as research instruments and need to be checked against human data or other ground truth and reported using appropriate performance metrics (e.g., accuracy, F1 score, precision/recall for classification; agreement measures such as Cohen's κ for annotation tasks). Just as survey questionnaires or interview guides are carefully piloted, documented, and archived, prompts should be systematically developed, tested, and reported in full. Publishing the exact prompt templates, including system instructions and formatting details, enables replication and critical evaluation. Usage of validation and cross-validation procedures would increase robustness, reliability, and potential bias in outputs.

Beyond these obvious concerns, researchers must also recognize less visible sources of variability. Interaction history, for example, can "contaminate" model outputs (e.g. Gupta et al., 2024). Many LLM interfaces maintain conversational context, meaning that earlier exchanges may subtly influence later responses. If prompts are tested interactively before being deployed in batch processing, residual context may affect results in ways that are difficult to detect. A clean session with explicitly controlled context can help to mitigate this. Automated scripts should reset or isolate interactions to ensure independence across observations.

For large language models inference parameters themselves are a critical component of reproducibility. Temperature and random seeds can all materially affect outputs. Deterministic settings, e.g. setting temperature set to zero with a fixed random seed are generally preferable for research applications requiring stability, although even so a full replication of answers is not necessarily guaranteed (e.g. Astekin et al., 2024). All such parameters should be explicitly fixed and documented in publications or supplementary materials.

Automation plays a central role in enforcing these standards. Rather than manually copying prompts into web interfaces, researchers should rely on scripts that programmatically send inputs and

capture outputs. This reduces human error, eliminates undocumented prompt drift, and creates a verifiable audit trail. Ideally, the entire pipeline from data preprocessing to model inference to post-processing and analysis should be encapsulated in reproducible scripts managed through version control systems. Logging raw outputs before cleaning or transformation is equally important, as it preserves the original generative record.

Thus, application of large language model should be regarded as similar to machine learning task and require openness of all data, code, experimental instructions, validation procedures and model settings to replicate the study findings.

It is worth noting that, that there could be other sources of variation as the same model may produce slightly different outputs depending on whether it runs on a CPU or GPU, or depending on the underlying architecture and driver stack. Precision settings and model quantization levels can further influence numerical stability and token selection, especially in borderline probability cases. While such differences may appear negligible, they can accumulate in large-scale coding tasks or classification pipelines. Consequently, researchers should standardize the computational environment as rigorously as possible and report hardware configurations, library versions, and inference settings.

Rigorous reporting of computational environment (like library versions) could help a lot in further replication, as well as containerization technologies, such as Docker or similar tools, provide an additional safeguard. By encapsulating the operating system, dependencies, and model files within a container, researchers can ensure that others can recreate the computational environment with minimal ambiguity. Locally run models within containers are particularly advantageous because they reduce reliance on changing external APIs and eliminate uncertainties about server-side updates.

Finally, transparency and reflexivity remain essential. LLMs are trained on vast and heterogeneous corpora, and their outputs may reflect embedded biases or normative assumptions. Social scientists must therefore critically evaluate model behavior, validate outputs against human-coded benchmarks where possible, and clearly communicate limitations. Reproducibility is not merely a technical requirement; it is a foundation for epistemic accountability.

In sum, integrating LLMs into social-science research demands the same rigor applied to any computational method augmented by heightened awareness of model volatility, prompt sensitivity, hidden variability, and infrastructural dependencies. Through careful documentation, parameter control, automation, environment standardization, and transparent reporting, researchers can harness the analytical potential of LLMs while upholding the standards of scientific integrity and reproducibility.

5. Toward reproducible AI-supported research

The use of large language models brings computational social science closer to computer science; thus, they can benefit from frameworks that support the reproducibility of computational experiments like (Costa et al., 2025) or other prediction models (e.g. Collins et al., 2024).

The growing integration of cloud-based tools in social science for data interpretation, classification, and modelling has amplified the importance of robust data management practices. First, reliance of on external infrastructures for processing behavioural or textual data may lead to the risk of personal and politically sensitive data leakage increases. Second, beyond direct breaches, more subtle unintended consequences may arise. For example, participants' voices, texts, or behavioural traces could be incorporated into model training pipelines or used for generative purposes without their explicit consent. Ethical data stewardship must therefore extend beyond anonymisation to include scrutiny of storage, processing agreements, secondary use policies, and long-term data governance.

At the same time, the growing complexity of contemporary models demands a lot of computational resources. As resource requirements escalate, the practical feasibility of reproducing results diminishes. This creates structural inequalities between well-funded institutions and smaller research groups, potentially undermining the collective credibility of scientific findings. Addressing this challenge requires coordinated collaboration and establishing of reproducibility networks.

We also see growing evidence that although LLMs perform well in some studies, despite their vastly greater complexity and resource requirements (e.g., Luccioni et al. 2024), they do not consistently outperform leaner, more efficient, and less resource-intensive predictive models (e.g., Zhou et al., 2024; Porshnev et al., 2024). Therefore, researchers should carefully consider the sustainability of their methodological choices and broader research practices.

We highlighted common mistakes that are present in the current landscape of AI-assisted research in psychology and social science discourses. Much research is conducted using inaccessible Large Language Models that come attached to an economic (Waelen & van Wynsberghe, 2025), ecological (Strubell et al., 2019), and/or ethical cost (Bommasani et al.,

2021), which reduces the responsibility and reproducibility of such scientific outputs. Further, singular black-box API outputs of cloud-hosted LLMs are often treated as stable observations (Morishige & Koshihara, 2025), ignoring the instability and inherent variability of these outputs. Beyond that, many studies fail to report the prompting, sampling, and fine-tuning parameters in full and thus provide only incomplete documentation of their methodology (Mitchell et al., 2019; Kapoor et al., 2024; Guo et al., 2025). We proposed that researchers should re-evaluate the need to use the computationally most powerful models for their research, since more efficient and resource intensive predictive models often lead to comparable results (Zhou et al., 2024; Porshnev et al., 2024). Using locally hosted open-access models would not only eliminate the issue of limited access hindering reproducibility but further allow for a more controlled and less instable environment to record LLM-outputs. Moreover, implementing a programmatic script to automatize and record all LLM-inputs and -outputs reduces variability and increases the stability of model outputs. Using such scripts and making them accessible together with all other replication materials further increases the reproducibility of a research output.

As AI becomes more prevalent in psychology and social science, researchers must become familiar with the frameworks and best practices of computational experiments (Costa et al., 2025; Collins et al., 2024; Abdurahman et al., 2025). The ongoing developments in large language model-assisted research require not only technical expertise, but also clear methodological guidelines and structured training programmes for scientists at all career stages. As technological capabilities evolve, further ethical, methodological, and infrastructural challenges are likely to emerge, demanding continuous reflection and proactive governance.

6. Conclusion & Future Work

Following the principles of open research practices that have strengthened scientific practice over the past decade, we outline a clear and accessible framework that improves researchers' awareness of LLM limitations and supports robust, transparent and reproducible knowledge when working with LLMs. In future work, we aim to conduct a systematic review of LLM-research across different disciplines within the social and cognitive sciences to further investigate reproducibility across different fields and to quantify which pitfalls occur in each area in what capacity. As Large Language Models continue to reshape the landscape of social science research, the responsibility to uphold reproducibility and integrity rests not only with individual researchers, but with the scientific community as a whole.

By embracing transparency, rigor, and interdisciplinary expertise beyond one's own discipline, grounded in the best practices outlined here, we can transform the challenges of AI-supported research into opportunities for more robust, trustworthy, and impactful science, ensuring that innovation never comes at the cost of credibility.

7. Copyrights

8. Acknowledgments

This publication has emanated from research conducted with the financial support of Science Foundation Ireland under Grant number 21/FFP-P/10118.

Bibliographical References

Abdurahman, S., Salkhordeh Ziabari, A., Moore, A. K., Bartels, D. M., & Dehghani, M. (2025). A Primer for Evaluating Large Language Models in Social-Science Research. Advances in Methods and Practices in Psychological Science, 8(2), 25152459251325174. https://doi.org/10.1177/25152459251325174

Ali K. Y., Akter S., Islam S., Mridha M. E. (2025). Advancing computational intelligence: AI-based algorithm design and optimization in programming. J. Comput. Sci. Technol. Stud. 7, 122–138. 10.32996/jcsts.2025.7.1.10

Artificial Analysis (2026, May 15) Artificial Analysis Openness Index Specification V1.0, https://artificialanalysis.ai/articles/announcing-artificial-analysis-openness-index

Astekin, M., Hort, M., & Moonen, L. (2024). An Exploratory Study on How Non-Determinism in Large Language Models Affects Log Parsing. Proceedings of the ACM/IEEE 2nd International Workshop on Interpretability, Robustness, and Benchmarking in Neural Software Engineering, 13–18. https://doi.org/10.1145/3643661.3643952

Atil, B., Chittams, A., Fu, L., Ture, F., Xu, L., Baldwin, B., … & Sidor, S. (2024). LLM Stability: A detailed analysis with some surprises.

Baker, M. (2016). 1,500 scientists lift the lid on

reproducibility. Nature, 533(7604), 452–454. https://doi.org/10.1038/533452a

Bakker, B. N., Jaidka, K., Dörr, T., Fasching, N., & Lelkes, Y. (2021). Questionable and open research practices: Attitudes and perceptions among quantitative communication researchers. Journal of Communication, 71(5), 715–738. https://doi.org/10.1093/joc/jqab031

Belz, A., Agarwal, S., Shimorina, A., & Reiter, E. (2021). A Systematic Review of Reproducibility Research in Natural Language Processing. Conference of the European Chapter of the Association for Computational Linguistics.

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021, March). On the dangers of stochastic parrots: Can language models be too big? 🦜. In Proceedings of the 2021 ACM conference on fairness, accountability, and transparency (pp. 610-623). https://doi.org/10.1145/3442188.3445922

Bhatia, S., & Walasek, L. (2023). Predicting implicit attitudes with natural language data. Proceedings of the National Academy of Sciences, 120(25), e2220726120. https://doi.org/10.1073/pnas.2220726120

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., ... & Liang, P. (2021). On the opportunities and risks of foundation models. arXiv preprint arXiv:2108.07258.

Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. Science, 356(6334), 183–186. https://doi.org/10.1126/science.aal4230

Chairs, P. (2023, January 10). ACL 2023 Policy on AI Writing Assistance. ACL 2023. https://acl-org.github.io/blog/ACL-2023-policy/

Chen, L., Zaharia, M., & Zou, J. (2024). How is ChatGPT's behavior changing over time?. Harvard Data Science Review, 6(2).

Christensen, G., & Miguel, E. (2018). Transparency, Reproducibility, and the Credibility of Economics Research. Journal of Economic Literature, 56(3), 920–980. https://doi.org/10.1257/jel.20171350

Collins, G. S., Moons, K. G. M., Dhiman, P., Riley, R. D., Beam, A. L., Van Calster, B., Ghassemi, M., Liu, X., Reitsma, J. B., Van Smeden, M., Boulesteix, A.-L., Camaradou, J. C., Celi, L. A., Denaxas, S., Denniston, A. K., Glocker, B., Golub, R. M., Harvey, H., Heinze, G., … Logullo, P. (2024). TRIPOD+AI statement: Updated guidance for reporting clinical prediction models that use regression or machine learning methods. BMJ, 385, e078378. https://doi.org/10.1136/bmj-2023-078378

Costa, L., Barbosa, S., & Cunha, J. (2025). A Framework for Supporting the Reproducibility of Computational Experiments in Multiple Scientific Domains (arXiv:2503.07080). arXiv. https://doi.org/10.48550/arXiv.2503.07080

Demszky, D., Yang, D., Yeager, D. S., Bryan, C. J., Clapper, M., Chandhok, S., ... & Pennebaker, J. W. (2023). Using large language models in psychology. Nature Reviews Psychology, 2(11), 688-701. https://doi.org/10.1038/s44159-023-00241-5

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

Fanelli, D. (2018). Is science really facing a reproducibility crisis, and do we need it to? Proceedings of the National Academy of Sciences, 115(11), 2628–2631. https://doi.org/10.1073/pnas.1708272114

Feng, S., Park, C. H., Liu, Y., Tsvetkov, Y. (2023). From Pretraining Data to Language Models to Downstream Tasks: Tracking the Trails of Political Biases Leading to Unfair NLP Models. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Vol 1), pages 11737-11762, Toronto, Canada.

Foster, E. D., & Deardorff, A. (2017). Open Science Framework (OSF). Journal of the Medical Library Association : JMLA, 105(2), 203–206. https://doi.org/10.5195/jmla.2017.88

Guo, Z., Lv, H., Zhang, C., Zhao, Y., Zhang, Y., & Cui, L. (2025, November). The Illusion of Randomness: How LLMs Fail to Emulate Stochastic Decision-Making in Rock-Paper-Scissors Games?. In Findings of the Association for Computational Linguistics: EMNLP 2025 (pp. 8618-8637). https://aclanthology.org/2025.findings-emnlp.458/

Gupta, A., Sheth, I., Raina, V., Gales, M., & Fritz, M. (2024). LLM Task Interference: An Initial Study on the Impact of Task-Switch in Conversational History. Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, 14633–14652. https://doi.org/10.18653/v1/2024.emnlp-main.811

Hartmann P., Henkel J. (2020). The rise of corporate science in AI: data as a strategic resource. Acad. Manag. Discov. 6, 359–381. 10.5465/amd.2019.004338778362

Hassija, V., Chamola, V., Mahapatra, A. et al. Interpreting Black-Box Models: A Review on Explainable Artificial Intelligence. Cogn Comput 16, 45–74 (2024). https://doi.org/10.1007/s12559-023-10179-8

Hutchinson, B., Prabhakaran, V., Denton, E., Webster, K., Zhong, Y., & Denuyl, S. (2020, July). Social biases in NLP models as barriers for persons with disabilities. In Proceedings of the 58th annual meeting of the association for computational linguistics (pp. 5491-5501).

Hutson, M. (2018). Artificial intelligence faces reproducibility crisis. https://doi.org/10.1126/science.359.6377.725

John, L. K., Loewenstein, G., & Prelec, D. (2012).

Measuring the Prevalence of Questionable Research Practices With Incentives for Truth Telling. Psychological Science, 23(5), 524–532. https://doi.org/10.1177/0956797611430953

Kapoor, S., Cantrell, E. M., Peng, K., Pham, T. H., Bail, C. A., Gundersen, O. E., ... & Narayanan, A. (2024). REFORMS: Consensus-based Recommendations for Machine-learning-based Science. Science Advances, 10(18), eadk3452. https://doi.org/10.1126/sciadv.adk3452

Kreutzer, J., Caswell, I., Wang, L., Wahab, A., van Esch, D., Ulzii-Orshikh, N., Tapo, A., Subramani, N., Sokolov, A., Sikasote, C., Setyawan, M., Sarin, S., Samb, S., Sagot, B., Rivera, C., Rios, A., Papadimitriou, I., Osei, S., Suarez, P. O., … Adeyemi, M. (2022). Quality at a glance: An audit of web-crawled multilingual datasets. arXiv.org. https://arxiv.org/abs/2103.12028

Liang, P., Bommasani, R., Lee, T., Tsipras, D., Soylu, D., Yasunaga, M., ... & Koreeda, Y. (2022). Holistic evaluation of language models. arXiv preprint arXiv:2211.09110. https://doi.org/10.48550/arXiv.2211.09110

Luccioni, S., Gamazaychikov, B., Hooker, S., Pierrard R., Strubell, E., Jernite, Y. , Wu, C.J. Light bulbs have energy ratings—so why can't AI chat-bots?, Nature 632 (2024) 736–738.

Lynott, D., Walsh, M., McEnery, T., Connell, L., Cross, L., & O'Brien, K. (2019). Are You What You Read? Predicting Implicit Attitudes to Immigration Based on Linguistic Distributional Cues From Newspaper Readership; A Pre-registered Study. Frontiers in Psychology, 10, 842. https://doi.org/10.3389/fpsyg.2019.00842

Magnusson, Ian & Smith, Noah & Dodge, Jesse. (2023). Reproducibility in NLP: What Have We Learned from the Checklist?. 10.48550/arXiv.2306.09562.

Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., ... & Gebru, T. (2019, January). Model cards for model reporting. In Proceedings of the conference on fairness, accountability, and transparency (pp. 220-229). https://doi.org/10.1145/3287560.3287596

Morishige, M., & Koshihara, R. (2025). Ensuring Reproducibility in Generative AI Systems for General Use Cases: A Framework for Regression Testing and Open Datasets. arXiv preprint arXiv:2505.02854. https://doi.org/10.48550/arXiv.2505.02854

Munafò, M. R., Chambers, C. D., Collins, A. M., Fortunato, L., & Macleod, M. R. (2020). Research culture and reproducibility. Trends in Cognitive Sciences, 24(2), 91–93. https://doi.org/10.1016/j.tics.2019.12.002

Munafò, M., Nosek, B., Bishop, D. et al. A manifesto for reproducible science. Nat Hum Behav 1, 0021 (2017). https://doi.org/10.1038/s41562-016-0021

Nelson, L. D., Simmons, J., & Simonsohn, U. (2018). Psychology's Renaissance. Annual Review of Psychology, 69, 511–534. https://doi.org/10.1146/annurev-psych-122216-011836

Niu, Q., Liu, J., Bi, Z., Feng, P., Peng, B., Chen, K., ... & Liu, M. (2024). Large language models and cognitive science: A comprehensive review of similarities, differences, and challenges. arXiv preprint arXiv:2409.02387.https://doi.org/10.48550/arXiv.2409.02387

Nosek, B. A., Alter, G., Banks, G. C., Borsboom, D., Bowman, S. D., Breckler, S. J., ... & Yarkoni, (2015). Promoting an open research culture. Science, 348(6242), 1422-1425. https://doi.org/10.1126/science.aab2374

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. Science, 349(6251), aac4716. https://doi.org/10.1126/science.aac4716

Paullada, A., Raji, I. D., Bender, E. M., Denton, E., & Hanna, A. (2021). Data and its (dis) contents: A survey of dataset development and use in machine learning research. Patterns, 2(11). https://doi.org/10.1016/j.patter.2021.100336

Pei, G., & Huang, H. (2025). Open science falling behind in the era of artificial intelligence. Frontiers in research metrics and analytics, 10, 1595824. https://doi.org/10.3389/frma.2025.1595824

Pineau, J., Vincent-Lamarre, P., Sinha, K., Larivière, V., Beygelzimer, A., d'Alché-Buc, F., ... & Larochelle, H. (2021). Improving reproducibility in machine learning research (a report from the neurips 2019 reproducibility program). Journal of machine learning research, 22(164), 1-20.

Porshnev, A., Kiy, K. D., O'Donoghue, D., Singh, M., Wingfield, C., & Lynott, D. (2024). Modeling Implicit Attitudes with Natural Language Data: A Comparison of Language Models. AICS 2024. The 32nd Irish Conference on Artificial Intelligence and Cognitive Science, Dublin, Ireland. https://aics2024.ucd.ie

Rettenberger, L., Reischl, M., Schutera, M. (2025). Assessing political bias in large language models. J Comput Soc Sc 8, 42 (2025). https://doi.org/10.1007/s42001-025-00376-w

Shiffrin, R., & Mitchell, M. (2023). Probing the psychology of AI models. Proceedings of the National Academy of Sciences, 120(10), e2300963120. https://doi.org/10.1073/pnas.2300963120

Smaldino, P. E., & McElreath, R. (2016). The natural selection of bad science. Royal Society Open Science, 3(9), 160384. https://doi.org/10.1098/rsos.160384

Strachan, J. W., Albergo, D., Borghini, G., Pansardi, O., Scaliti, E., Gupta, S., ... & Becchio, C. (2024). Testing theory of mind in large language models and humans. Nature

Human Behaviour, 8(7), 1285-1295. https://www.nature.com/articles/s41562-024-01882-z

Strubell, E., Ganesh, A., & McCallum, A. (2019, July). Energy and policy considerations for deep learning in NLP. In Proceedings of the 57th annual meeting of the association for computational linguistics (pp. 3645-3650). https://doi.org/10.48550/arXiv.1906.02243

Waelen, R., & van Wynsberghe, A. (2025). Considering the Social and Economic Sustainability of AI. Science and engineering ethics, 31(4), 19. https://doi.org/10.1007/s11948-025-00544-1

Walker, V., Angst, M. (2025). Promises and pitfalls of using LLMs to identify actor stances in political discourse. PloS one, 20(11), e0335547. https://doi.org/10.1371/journal.pone.0335547

Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., ... & Fedus, W. (2022). Emergent abilities of large language models. arXiv preprint arXiv:2206.07682.

Zhou, Y., Xu, P., Liu, X., An, B., Ai, W., Huang, F. (2024). Explore Spurious Correlations at the Concept Level in Language Models for Text Classification. http://arxiv.org/abs/2311.08648 (accessed September 19, 2024).

9. Language Resource References

All of our scientific outputs, data, and analysis can be found here: https://osf.io/gv2u7/