

A reporting checklist for large language models in behavioural science

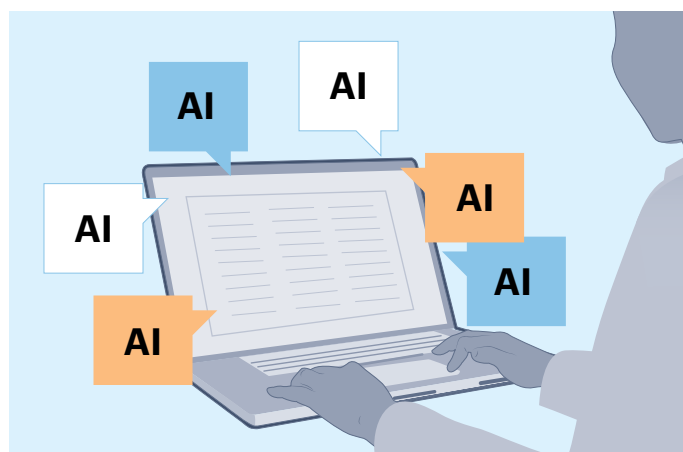
Stefan Feuerriegel, Christopher Barrie, M. J. Crockett, Laura K. Globig, Killian L. McLoughlin, Dan-Mircea Mirea, Arthur Spirling, Diyi Yang, Tim Althoff, Maria Antoniak, Lisa P. Argyle, Ashwini Ashokkumar, Mohammad Atari, Hannah Bailey, Kevin Bauer, Umang Bhatt, Yidong Chai, Tanmoy Chakraborty, Yanto Chandra, Huimin Chen, Hal Daumé III, Gianmarco De Francisci Morales, Morteza Dehghani, Danica Dillion, Johannes C. Eichstaedt, Kerstin Forster, Dominique Geissler, Kurt Gray, Thomas L. Griffiths, Hartmann Jochen, Oliver P. Hauser, James K. He, Rahul Hemrajani, Felix Holzmeister, Angel Hsing-Chi Hwang, Tiancheng Hu, Anna A. Ivanova, Nils Köbis, Yara Kyrychenko, Himabindu Lakkaraju, Jia Liu, Abdurahman Maarouf, Sebastian Maier, Lennart Meincke, Rada Mihalcea, Brent Mittelstadt, Saif M. Mohammad, Mor Naaman, Oded Netzer, Alice Oh, Desmond C. Ong, Francesco Pierri, Barbara Plank, Iyad Rahwan, Talal Rahwan, Pooja S. B. Rao, Claire E. Robertson, David M. Rothschild, Matthew J. Salganik, Eric Schulz, Chirag Shah, Yash Raj Shrestha, Ekaterina Shutova, Alexandra A. Siegel, Almog Simchon, Huan Sun, Malte Toetzke, Jay J. Van Bavel, Michelle Vaccaro, Jennifer Wortman Vaughan, Effy Vayena, Pedro O. S. Vaz-de-Melo, Briana Vecchione, Angelina Wang, Robert West, Robb Willer, Dirk U. Wulff, Renwen Zhang, Simone Zhang, Steve Rathje & Manoel Horta Ribeiro



Large language models offer new opportunities for behavioural science, but their rapid evolution poses challenges for research rigour. We introduce a consensus-based reporting checklist to improve transparency, reproducibility and ethical accountability of large-language-model-based research in the behavioural sciences.

Large language models (LLMs) are deep neural network architectures (typically transformers) trained on a large body of textual data that can generate human-like text. Many researchers across the behavioural and social sciences are enthusiastic about the potential of LLMs to open new avenues for studying human behaviour¹. For example, researchers have proposed using LLMs to conduct *in silico* experiments that simulate human judgments and decisions in response to interventions, and to facilitate large-scale data annotation and analysis². Other works have deployed LLMs as interventions to foster creativity, persuade, teach, or reduce misinformation beliefs^{3,4}. However, the rapidly evolving role of LLMs in shaping empirical evidence, theoretical frameworks, policy decisions, and public discourse poses challenges for research rigour.

Here, we present a consensus-based [checklist](#) ('GUIDE-LLM') for research that involves LLMs in behavioural and social science, with the aim of strengthening transparency, reproducibility and ethical accountability. GUIDE-LLM stands for 'guidelines for the use of LLMs in behavioural and social science'. The GUIDE-LLM checklist provides researchers with concrete guidance on improving transparency, reproducibility and ethical use, to strengthen rigour and documentation



throughout the entire research workflow (including when LLMs function as research tools, as well as studies in which LLMs themselves are the object of empirical investigation).

Challenges of LLM use

The way in which LLMs are implemented, used and reported in scientific research can vary widely. For example, the label 'ChatGPT' can refer to different underlying models (for example, GPT-4 or GPT-4o), each with multiple versions that are often marked by timestamps (for example, gpt-4o-2024-11-20). Other LLMs (such as Llama) are available in different sizes even within the same version number (for example, 8 billion versus 70 billion parameters), which leads to large performance differences across tasks. Even with the same model, behaviour may differ depending on the access mode (for example, via the official API (application programming interface) or a web interface), due to differences

in system prompts (specific prompts with predefined instructions before any user input to control the model's behaviour) and runtime environments (for example, differences in safety layers, inference infrastructure or memory use). These sources of heterogeneity can lead to replication failures, especially for commercial models whose developers are not obligated to publicly disclose model changes. This challenge is compounded by the fact that access is often controlled by commercial companies and that access may be modified or discontinued without notice, which makes long-term reproducibility and verification difficult⁵.

Outputs from LLMs may also vary as a product of specific parameter settings, and small differences could yield different outcomes even for identical prompts. For instance, the 'temperature' parameter controls the randomness of responses; higher values generate more diverse outputs across runs, and lower values generate more deterministic outputs. Such randomness can substantially influence behaviour, especially in tasks that involve text generation or reasoning. Even with a temperature set to zero, outputs can still vary owing to hardware-level non-determinism (note that low temperatures are not necessarily preferable but depend on the research objective, such as benefits from greater exploration versus more consistent responses). Further, the 'token limit' can be used as a parameter to control the length of the output, which may affect performance. Finally, we note that some frontier models (for example, GPT-5) no longer allow parameters such as temperature to be set at all, which illustrates that even the set of parameters available for reproducibility can change over time.

Prompts are another major source of variability⁶. Slight differences in how prompts are phrased can markedly change model outputs. One can steer an LLM's behaviour through various techniques such as personas (instructing the model to respond from a specific perspective or role), chain-of-thought reasoning (guiding the model to solve problems step by step) or in-context learning (providing examples within the prompt). But differences in prompts are often unrecorded and/or unreported, which makes it difficult to compare results across studies.

The transparency of LLMs is further complicated by the fact that training data are typically hidden from end-users. Because LLMs are trained on vast datasets that may contain biased, incomplete or sensitive information, they can reproduce existing societal biases and inequities^{7,8}, including stigmatizing and stereotypical language. For downstream research using LLMs in behavioural or social science, this upstream opacity remains a salient limitation, so a key challenge is to recognize and document how such model characteristics may shape one's own study designs and conclusions. Moreover, LLMs may have been exposed to specific study stimuli or benchmark materials during training, which complicates interpretation and threatens internal validity (and memorization tests are often imperfect). Without careful documentation, validation and safeguards, studies in behavioural and social science may overlook issues such as the propagation of societal biases or data contamination (for example, prior exposure of LLMs to test materials during training), and thereby complicates the interpretation and robustness of findings that involve LLMs. Researchers must also account for LLMs with memory, in which prior exposure to study materials in previous sessions may influence subsequent outputs (which makes reproducibility challenging).

Consensus method

The checklist was developed through a preregistered, two-round, reactive Delphi study⁹, with input from an expert panel of researchers experienced in conducting LLM-based studies across various

subdisciplines from behavioural and social sciences (for example, psychology, political science, economics and sociology) as well as machine-learning researchers specialized in natural language processing. The method is described in our supplementary materials¹⁰. In round 1, an expert panel ($n = 68$) evaluated an initial pool of candidate items using a five-point Likert-type scale that ranged from 'strongly exclude' to 'strongly include'. Candidate items with positive scores were provisionally retained and revised after careful discussion among the core members to accommodate the qualitative feedback. No additional items were added during round 1, which reflects the fact that the initial item pool had been deliberately designed to be comprehensive by covering the broad range of potential reporting dimensions identified from prior literature, related reporting frameworks and prior experience. In round 2, the refined checklist was evaluated by the experts ($n = 80$, which included the experts from round 1 as well as additional experts to improve geographical representation), but now with a binary inclusion vote ('include' versus 'exclude'). Details are available in our supplementary materials¹⁰.

We used a preregistered consensus threshold of more than two-thirds of votes in favour of inclusion, so that all items reached broad expert support for their inclusion. Items that did not meet this threshold were removed from the checklist, which implies that the resulting items in the checklist reflect broad agreement among experts. Several items that did not reach the inclusion threshold were eventually retained as optional items in the online version of the checklist.

Our consensus-based approach has limitations, including overlap between the author group and expert panel, uneven geographical representation, and the possibility that some context-specific but important items did not meet the predefined inclusion threshold; further details are provided in our supplementary materials¹⁰.

The GUIDE-LLM checklist

We present a consensus-based checklist that behavioural and social science researchers can use to improve and document the transparency of their research (Table 1). Overall, the GUIDE-LLM checklist comprises 14 items, which cover the scope of LLM use (2 items), model or system details (5 items), prompts (2 items), data inputs and privacy (1 item), validation and interpretation (2 items), reproducibility (1 item) and competing interests (1 item). The majority of items are broadly applicable across research designs, and hence the checklist should serve as a minimum reporting standard. Below, we highlight key elements of the checklist.

The checklist emphasizes that rigorous reporting begins with clarity about how and why LLMs are used. In behavioural and social sciences, LLMs can appear at nearly any stage of the research process – from designing stimuli and coding qualitative data to serving as participants in simulated experiments or as interventions interacting with humans. Each of these roles raises distinct methodological and ethical considerations. For instance, an LLM used to generate experimental materials requires clear documentation of prompt design and validation, and an LLM used as a participant-facing chatbot calls for particular attention to safety, oversight and informed consent.

The checklist highlights the importance of exact reporting about the model or system (for example, often expressed as a timestamp, such as 'gpt-4o-2024-11-20'), and of including the exact prompts used. Many published studies refer broadly to 'ChatGPT' or 'GPT-4', yet these broad labels mask substantial variation across versions, configurations and access modes that can materially affect results. Further, validation of LLM outputs is important because LLMs are often prompted to perform

Table 1 | The GUIDE-LLM checklist for promoting transparency, reproducibility and ethical accountability in studies that involve LLMs

Item	Name
Scope of LLM use	
Item A.1	LLMs were used in this project for: ...
Item A.2	Degree of automation (human in the loop versus fully automated): ...
Model or system details	
Item B.1	Model name, including provider, model size, exact version or identifier, date of access and source link (if possible): ...
Item B.2	Model access (for example, API, web interface or local) and context mode (for example, chat mode or separate calls): ...
Item B.3	Relevant LLM configuration(s) reported (as applicable), such as temperature, maximum tokens, seed and number of runs: ...
Item B.4	Customization: ...
Item B.5	Did the LLM session(s) include persistent memory across interactions? ...
Prompts	
Item C.1	Exact prompt(s) reported: ...
Item C.2	System-wide instructions (if any): ...
Data inputs and privacy	
Item D.1	Handling of personal or sensitive data (if any) (for example, consent for data processing): ...
Validation and interpretation	
Item E.1	Human validation of LLM outputs: ...
Item E.2	Describe any relevant post-processing (for example, filtering in case of format mismatches, unit conversions and so on): ...
Reproducibility	
Item F.1	Code, notebooks or scripts for LLM calls shared: ...
Competing interests	
Item G.1	Funding, support or other relevant relationships (including in-kind access to compute or models, or professional affiliations): ...

A template for the GUIDE-LLM checklist is also available [online](#). The answer options are omitted from the above table and provided in the template.

specific analytic tasks such as identifying sentiment, detecting emotions or classifying moral language, so human validation helps to verify that the LLM’s responses indeed reflect the intended construct². The quality of this validation directly affects how confidently researchers can interpret downstream findings and assess the reliability of LLM-assisted studies. Finally, the checklist emphasizes the importance of sharing code, scripts and example interactions – while carefully redacting any sensitive information – to enable other researchers to verify findings and adapt methods to new contexts. Such openness is particularly crucial in a fast-moving field in which LLM versions and access conditions may change or become discontinued. Moreover, because many researchers using LLMs may have financial or professional ties to major technology companies or have received benefits through computing resources¹¹, the checklist urges transparent disclosure of any potential competing interests. Clear reporting of such relationships helps readers to assess possible sources of bias and ultimately strengthens trust in LLM-based behavioural science research.

The online version of the GUIDE-LLM checklist also includes a list of optional items that did not reach consensus during the Delphi process but were nonetheless considered valuable by many experts. The optional items focus on: (1) the justification for the LLM choice; (2) the rationale for the prompt design; (3) comparison against other LLMs or methods (for example, whether the study conclusions are unique to a specific LLM or are generalizable across different LLMs); (4) the risks of training data leakage (for example, to reflect on cases in which LLMs may have been exposed to test materials during training, potentially conflating performance estimates); (5) assessment of potential bias or systematic differences in LLM behaviour that could affect the study’s conclusions (for example, LLMs are known to display gender, racial or cultural bias that could affect results); (6) conversation transcripts; (7) a discussion of ethical implications (for example, such as when LLM chatbots are used for participant-facing interventions); and (8) computational resource use.

Practical considerations

The checklist is intended to serve a wide range of stakeholders: behavioural and social science researchers authoring LLM studies; journal editors, peer reviewers and scientific readers evaluating research; policy-makers seeking to implement rules and regulations; and the broader public that benefits from increased transparency and rigour of LLM-based research. The GUIDE-LLM checklist is not intended to prescribe specific modelling choices for LLM-based research, but rather to promote transparency.

Although the majority of items should be broadly applicable, some flexibility is necessary. We therefore advocate a pragmatic, context-sensitive use of the checklist. In such cases, researchers may leave items blank or may explain why they are inapplicable. Further, the checklist is aimed at cases in which LLM use is integral to the research design rather than minor or purely editorial uses of LLMs. When multiple LLMs are used for different tasks or steps in the research process, the relevant sections should be completed separately for each model. We also recognize pragmatic constraints: for example, in agentic systems, reporting verbatim prompts may be impractical, in which case researchers should provide the most complete documentation feasible (for example, source code, configuration files or structured descriptions of the interaction logic). To simplify reporting, authors are encouraged to reference the relevant sections, pages or appendices of their manuscripts rather than duplicating text.

The checklist is maintained as a ‘living document’ to allow for regular updates. Further, the GUIDE-LLM was developed primarily with text-only models in mind, but most reporting items are applicable to multimodal LLMs and artificial intelligence agents. Reporting parameters such as temperature, seeds or other configuration settings does not guarantee full reproducibility but is rather a necessary first step towards greater transparency.

Stefan Feuerriegel ^{1,2}✉, Christopher Barrie ³, M. J. Crockett^{4,5}, Laura K. Globig ⁶, Killian L. McLoughlin⁷, Dan-Mircea Mirea ⁴, Arthur Spirling⁷, Diyi Yang^{8,9}, Tim Althoff ¹⁰, Maria Antoniak¹¹, Lisa P. Argyle ¹², Ashwini Ashokkumar¹³, Mohammad Atari ¹⁴, Hannah Bailey¹⁵, Kevin Bauer¹⁶, Umang Bhatt ¹⁷, Yidong Chai¹⁸, Tanmoy Chakraborty^{19,20}, Yanto Chandra²¹, Huimin Chen²², Hal Daumé III²³, Gianmarco De Francisci Morales²⁴, Morteza Dehghani ²⁵, Danica Dillion ²⁶, Johannes C. Eichstaedt ^{27,28}, Kerstin Forster^{1,2}

Dominique Geissler^{1,2}, **Kurt Gray**²⁹, **Thomas L. Griffiths**^{10,4,30}, **Hartmann Jochen**³¹, **Oliver P. Hauser**^{32,33}, **James K. He**³⁴, **Rahul Hemrajani**³⁵, **Felix Holzmeister**³⁶, **Angel Hsing-Chi Hwang**²⁵, **Tiancheng Hu**¹⁷, **Anna A. Ivanova**³⁷, **Nils Köbis**³⁸, **Yara Kyrychenko**¹⁷, **Himabindu Lakkaraju**³⁹, **Jia Liu**⁴⁰, **Abdurahman Maarouf**^{1,2}, **Sebastian Maier**^{1,2}, **Lennart Meincke**^{41,42}, **Rada Mihalcea**⁴³, **Brent Mittelstadt**^{44,45}, **Saif M. Mohammad**⁴⁶, **Mor Naaman**⁴⁷, **Oded Netzer**⁴⁸, **Alice Oh**⁴⁹, **Desmond C. Ong**⁵⁰, **Francesco Pierri**⁵¹, **Barbara Plank**¹, **Iyad Rahwan**⁵², **Talal Rahwan**⁵³, **Pooja S. B. Rao**⁵⁴, **Claire E. Robertson**⁵⁵, **David M. Rothschild**⁵⁶, **Matthew J. Salganik**⁵⁷, **Eric Schulz**⁵⁸, **Chirag Shah**⁵⁹, **Yash Raj Shrestha**⁶⁰, **Ekaterina Shutova**⁶¹, **Alexandra A. Siegel**¹¹, **Almog Simchon**⁶², **Huan Sun**²⁹, **Malte Toetzkke**^{63,64}, **Jay J. Van Bavel**⁶⁵, **Michelle Vaccaro**⁶⁶, **Jennifer Wortman Vaughan**⁶⁶, **Effy Vayena**⁶⁷, **Pedro O. S. Vaz-de-Melo**⁶⁸, **Briana Vecchione**⁶⁹, **Angelina Wang**⁴⁷, **Robert West**⁷⁰, **Robb Willer**⁷¹, **Dirk U. Wulff**^{72,73}, **Renwen Zhang**⁷⁴, **Simone Zhang**³, **Steve Rathje**³ & **Manoel Horta Ribeiro**⁷⁵

¹LMU Munich, Munich, Germany. ²Munich Center for Machine Learning, Munich, Germany. ³New York University, New York, NY, USA. ⁴Department of Psychology, Princeton University, Princeton, NJ, USA. ⁵University Center for Human Values, Princeton University, Princeton, NJ, USA. ⁶Department of Psychology and Neural Science, New York University, New York, NY, USA. ⁷Princeton University, Princeton, NJ, USA. ⁸Stanford NLP Group, Stanford University, Stanford, CA, USA. ⁹Stanford Institute for Human-Centered Artificial Intelligence (HAI), Stanford University, Stanford, CA, USA. ¹⁰Allen School of Computer Science & Engineering, University of Washington, Seattle, WA, USA. ¹¹University of Colorado Boulder, Boulder, CO, USA. ¹²Purdue University, West Lafayette, IN, USA. ¹³Department of Psychology, Harvard University, Cambridge, MA, USA. ¹⁴Department of Psychological and Brain Sciences, University of Massachusetts Amherst, Amherst, MA, USA. ¹⁵Carnegie Mellon Institute for Strategy and Technology, Carnegie Mellon University, Pittsburgh, PA, USA. ¹⁶Goethe University Frankfurt, Frankfurt, Germany. ¹⁷University of Cambridge, Cambridge, UK. ¹⁸City University of Hong Kong, Hong Kong, China. ¹⁹Indian Institute of Technology Delhi, New Delhi, India. ²⁰Indian Institute of Technology Delhi, Abu Dhabi, UAE. ²¹Department of Public and International Affairs, City University of Hong Kong, Hong Kong, China. ²²School of Journalism and Communication, Tsinghua University, Beijing, China. ²³Department of Computer Science and AI Interdisciplinary Institute at Maryland, University of Maryland-College Park, College Park, MD, USA. ²⁴CENTA, Turin, Italy. ²⁵University of Southern California, Los Angeles, CA, USA. ²⁶Complexity Science Hub; The Ohio State University, Columbus, OH, USA. ²⁷Department of Psychology, Stanford University, Stanford, CA, USA. ²⁸Decision Sciences, INSEAD, Paris, France. ²⁹The Ohio State University, Columbus, OH, USA. ³⁰Department of Computer Science, Princeton University, Princeton, NJ, USA. ³¹TUM School of Management, Technical University of Munich, Munich, Germany. ³²Department of Economics, University of Exeter, Exeter, UK. ³³Institute for Data Science and Artificial Intelligence, University of Exeter, Exeter, UK. ³⁴Societies, Inc, New York, NY, USA. ³⁵National Law School of India University, Bengaluru, India. ³⁶Department of Economics, University of Innsbruck, Innsbruck, Austria. ³⁷School of Psychology, Georgia Institute of Technology, Atlanta, GA, USA. ³⁸Research Center Trustworthy Data Science and Security, University Duisburg-Essen, Duisburg, Germany. ³⁹Harvard University, Cambridge,

MA, USA. ⁴⁰Hong Kong University of Science and Technology, Hong Kong, China. ⁴¹WHU - Otto Beisheim School of Management, Vallendar, Germany. ⁴²Mack Institute for Innovation Management, The Wharton School, University of Pennsylvania, Philadelphia, PA, USA. ⁴³University of Michigan, Ann Arbor, MI, USA. ⁴⁴Oxford Internet Institute, University of Oxford, Oxford, UK. ⁴⁵Weizenbaum Institute, Berlin, Germany. ⁴⁶National Research Council, Ottawa, Ontario, Canada. ⁴⁷Cornell Tech, New York, NY, USA. ⁴⁸Columbia University, New York, NY, USA. ⁴⁹KAIST, Daejeon, South Korea. ⁵⁰The University of Texas at Austin, Austin, TX, USA. ⁵¹Politecnico di Milano, Milan, Italy. ⁵²Max Planck Institute for Human Development, Center for Humans & Machines, Berlin, Germany. ⁵³New York University Abu Dhabi, Abu Dhabi, United Arab Emirates. ⁵⁴International Institute of Information Technology Bangalore, Bengaluru, India. ⁵⁵Colby College, Waterville, ME, USA. ⁵⁶Microsoft Research, New York, NY, USA. ⁵⁷Department of Sociology, Princeton University, Princeton, NJ, USA. ⁵⁸Institute for Human-Centered AI, Helmholtz Computational Health Center, Munich, Germany. ⁵⁹University of Washington, Seattle, WA, USA. ⁶⁰University of Lausanne, Lausanne, Switzerland. ⁶¹University of Amsterdam, Amsterdam, Netherlands. ⁶²Department of Psychology, Ben-Gurion University of the Negev, Beersheba, Israel. ⁶³Technical University of Munich, Munich, Germany. ⁶⁴Max Planck Institute for Innovation and Competition, Munich, Germany. ⁶⁵Norwegian School of Economics, Bergen, Norway. ⁶⁶Institute for Data, Systems, and Society (IDSS), MIT, Cambridge, MA, USA. ⁶⁷Health Ethics and Policy Lab, ETH Zurich, Zurich, Switzerland. ⁶⁸Universidade Federal de Minas Gerais, Belo Horizonte, Brazil. ⁶⁹Data & Society Research Institute, New York, NY, USA. ⁷⁰EPFL, Lausanne, Switzerland. ⁷¹Stanford University, Stanford, CA, USA. ⁷²Center for Adaptive Rationality, Max Planck Institute for Human Development, Berlin, Germany. ⁷³Faculty of Psychology, University of Basel, Basel, Switzerland. ⁷⁴Nanyang Technological University, Singapore, Singapore. ⁷⁵Department of Computer Science, Princeton University, Princeton, NJ, USA.

✉ e-mail: feuerriegel@lmu.de

Published online: 9 June 2026

References

- Messeri, L. & Crockett, M. J. *Nature* **627**, 49–58 (2024).
- Feuerriegel, S. et al. *Nat. Rev. Psychol.* **4**, 96–111 (2025).
- Argyle, L. P. et al. *Proc. Natl Acad. Sci. USA* **120**, e2311627120 (2023).
- Costello, T. H., Pennycook, G. & Rand, D. G. *Science* **385**, eadq1814 (2024).
- Palmer, A., Smith, N. A. & Spirling, A. *Nat. Comput. Sci.* **4**, 2–3 (2024).
- Yang, C. et al. In *International Conference on Learning Representations* <https://openreview.net/forum?id=Bb4VGOWELI> (2023).
- Bender, E. M., Gebru, T., McMillan-Major, A. & Shmitchell, S. On the dangers of stochastic parrots: can language models be too big? In *FAccT '21: Proc. 2021 ACM Conf. on Fairness, Accountability, and Transparency*, 610–623 (ACM, 2021).
- Hu, T. et al. *Nat. Comput. Sci.* **5**, 65–75 (2025).
- Grime, M. M. & Wright, G. in *Wiley StatsRef: Statistics Reference Online* <https://doi.org/10.1002/9781118445112.stat07879> (Wiley, 2016).
- Supplementary Materials.pdf. OSF <https://osf.io/mv63j/files/t2vwb> (2026).
- Ahmed, N., Wahed, M. & Thompson, N. C. *Science* **379**, 884–886 (2023).

Author contributions

The core team members (C.B., M.J.C., S.F., L.K.G., K.L.M., D.-M.M., M.H.R., S.R., A. Spirling and D.Y.) prepared the first version of the checklist items and reviewed the feedback from the Delphi study. S.F. conducted the survey study, analysed the data and drafted the initial version of the manuscript, with feedback from the core team members. All authors participated as experts in the Delphi survey and reviewed and edited the manuscript.

Funding

Funding from the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under the National Research Data Infrastructure – NFDI 27/1-2026, project number 460037581

is acknowledged. M.J.C. was supported by a grant from the Princeton Catalysis Initiative. T.A. was supported in part by NSF IIS-1901386, NSF CAREER IIS-2142794, and NIH R01MH125179. Y. Chai was supported by NSFC (72188101, 72342011). Y. Chandra was supported by a grant from the Research Grants Council of the Hong Kong Special Administrative Region, China (Project No. CityU 15513723). O.P.H. is grateful to the UKRI Future Leaders Fellowship (project reference: 1067). H.L. was supported in part by NSF awards IIS-2008461, IIS-2040989, and IIS-2238714; and an AI2050 Early Career Fellowship from Schmidt Sciences. R.M. was supported in part by NSF IIS-2306372. T.C. was supported by the Rajiv Khemani Young Faculty Chair Professorship in AI. M.J.S. was supported by grants from Princeton AI Lab, Princeton Precision Health and Princeton Catalysis Initiative. B.V. was supported in part by NSF IIS-2433429 and Internet Society Foundation grant 202503-22063. D.C.O. was supported in part by NSF CAREER HCC-2443038.

Competing interests

T.A. has received funding from Google, Microsoft and Apple, and is a part-time staff research scientist at Google. C.B. has received funding from OpenAI and Google. H.D. has received funding from Amazon, Google, Meta and the Schmidt Foundation, and was previously employed at Microsoft Research. M.D. has received funding from Google and Microsoft, and is a part-time senior research scientist (contractor) at Apple. J.C.E. has received consulting income and equity from Sonar Health and Jimini Health. L.K.G. has received funding from Google. T.L.G. has received funding from OpenAI, Google and Meta. O.P.H. has collaborated with Google Deepmind and received funding from Anthropic, UKRI and the UK AI Security

Institute. T.H. has received funding from Apple, Google and OpenAI. A.A.I. has received funding from Ferguson Control Systems (via NSF BRAIN). H.L. is a part-time senior staff research scientist at Google. This work was conducted entirely in her capacity as a faculty member at Harvard University. She has also received research funding from Google and OpenAI. O.N. is an Amazon scholar working in areas unrelated to this work. S.R. and J.J.V.B. have received research funding from OpenAI and Google. E. Shutova has received funding from Google, Meta and Deloitte. H.S. has received funding from Amazon, Google, Cisco, OpenAI, Fujitsu and Intuit. R.M. has received funding from OpenAI and LG AI, and in-kind support from Amazon and Microsoft. A. Simchon has received funding from Google. M.H.R. has received funding from Meta and OpenAI. A.O. received funding from Google, Microsoft, LG AI, and Naver. A.O. is a faculty visiting researcher at Google, but this work was done in her capacity as a faculty member at KAIST. D.C.O. has received research funding from Microsoft Research, Toyota Research Institute and OpenAI. B.V. has received funding from Microsoft and the Notre Dame-IBM Tech Ethics Lab unrelated to this work. M.J.S. has previously received funding from Facebook/Meta and Google, he was previously a consultant for Google, and he was previously employed at Microsoft Research. The other authors declare no competing interests.

Additional information

Peer review information *Nature Human Behaviour* thanks Zhicheng Lin, Thomas Pfeiffer and Robert Thibault for their contribution to the peer review of this work.