A consensus-based reporting checklist for large language models in behavioral and social science

Stefan Feuerriegel, Munich Center for Machine Learning & LMU Munich, Munich, Germany, feuerriegel@lmu.de

Christopher Barrie, New York University, cb5691@nyu.edu

M.J. Crockett, Department of Psychology, Princeton University & University Center for Human Values, Princeton University, mj.crockett@princeton.edu

Laura K. Globig, Department of Psychology and Neural Science, New York University, laura.globig@gmail.com

Killian L McLoughlin, Princeton University, k.mcloughlin@princeton.edu

Dan-Mircea Mirea, Department of Psychology, Princeton University, dmirea@princeton.edu

Arthur Spirling, Princeton University, arthur.spirling@princeton.edu

Diyi Yang, Stanford NLP Group & Stanford Institute for Human-Centered Artificial Intelligence (HAI), Stanford University, diyiy@stanford.edu

Tim Althoff, Allen School of Computer Science & Engineering, University of Washington, althoff@cs.washington.edu

Maria Antoniak, University of Colorado Boulder, maria.antoniak@colorado.edu

Lisa P. Argyle, Purdue University, largyle@purdue.edu

Ashwini Ashokkumar, Department of Psychology, Harvard University, ashwiniashokkumar@g.harvard.edu

Mohammad Atari, Department of Psychological and Brain Sciences, University of Massachusetts Amherst, matari@umass.edu

Hannah Bailey, Carnegie Mellon Institute for Strategy and Technology, Carnegie Mellon University, hbailey@andrew.cmu.edu

Kevin Bauer, Goethe University Frankfurt, bauer@wiwi.uni-frankfurt.de

Umang Bhatt, University of Cambridge, usb20@cam.ac.uk

Yidong Chai, City University of Hong Kong, yidong.chai@cityu.edu.hk

Tanmoy Chakraborty, Indian Institute of Technology Delhi, India, tanchak@iitd.ac.in

Yanto Chandra, Department of Public and International Affairs, City University of Hong Kong, ychandra@cityu.edu.hk

Huimin Chen, School of Journalism and Communication, Tsinghua University, huimchen@tsinghua.edu.cn

Hal Daumé III, Department of Computer Science and AI Interdisciplinary Institute at Maryland, University of Maryland-College Park, hal3@umd.edu

Gianmarco De Francisci Morales, CENTAI, gdfm@acm.org

Morteza Dehghani, University of Southern California, mdehghan@usc.edu

Danica Dillion, Complexity Science Hub; The Ohio State University, danicajdillion@gmail.com

Johannes C. Eichstaedt, Department of Psychology, Stanford University; Decision Sciences, INSEAD, johannes.stanford@gmail.com

Kerstin Forster, Munich Center for Machine Learning & LMU Munich, kerstin.forster@lmu.de

Dominique Geissler, Munich Center for Machine Learning & LMU Munich, d.geissler@lmu.de

Kurt Gray, The Ohio State University, kurtjgray@gmail.com

Thomas L. Griffiths, Departments of Psychology and Computer Science, Princeton University, Tomg@princeton.edu

Jochen Hartmann, TUM School of Management, Technical University of Munich, jochen.hartmann@tum.de

Oliver P. Hauser, Department of Economics and Institute for Data Science and Artificial Intelligence, University of Exeter, o.hauser@exeter.ac.uk

James K. He, Societies, Inc., james@societies.io

Rahul Hemrajani, National Law School of India University, Bengaluru, India, rahul.hemrajani@nls.ac.in

Felix Holzmeister, Department of Economics, University of Innsbruck, Innsbruck, Austria, felix.holzmeister@uibk.ac.at

Angel Hsing-Chi Hwang, University of Southern California, angel.hwang@usc.edu

Tiancheng Hu, University of Cambridge, th656@cam.ac.uk

Anna A. Ivanova, School of Psychology, Georgia Institute of Technology, a.ivanova@gatech.edu

Nils Köbis, Research Center Trustworthy Data Science and Security, University Duisburg-Essen, Nils.koebis@uni-due.de

Yara Kyrychenko, University of Cambridge, yk408@cam.ac.uk

Himabindu Lakkaraju, Harvard University, hlakkaraju@hbs.edu

Jia Liu, Hong Kong University of Science and Technology, jialiu@ust.hk

Abdurahman Maarouf, Munich Center for Machine Learning & LMU Munich, a.maarouf@lmu.de

Sebastian Maier, Munich Center for Machine Learning & LMU Munich, maier.sebastian@campus.lmu.de

Lennart Meincke, WHU - Otto Beisheim School of Management; Mack Institute for Innovation Management & The Wharton School, University of Pennsylvania, lennart@sas.upenn.edu

Rada Mihalcea, University of Michigan, mihalcea@umich.edu

Brent Mittelstadt, Oxford Internet Institute, University of Oxford & Weizenbaum Institute, brent.mittelstadt@oii.ox.ac.uk

Saif M. Mohammad, National Research Council Canada, saif.mohammad@nrc-cnrc.gc.ca

Mor Naaman, Cornell Tech, mor.naaman@cornell.edu

Oded Netzer, Columbia University, onetzer@gsb.columbia.edu

Alice Oh, KAIST, alice.oh@kaist.edu

Desmond C. Ong, The University of Texas at Austin, desmond.c.ong@gmail.com

Barbara Plank, LMU Munich, b.plank@lmu.de

Francesco Pierri, Politecnico di Milano, francesco.pierri@polimi.it

Iyad Rahwan, Max Planck Institute for Human Development, Center for Humans & Machines, Berlin, Germany, Rahwan@mpib-berlin.mpg.de

Talal Rahwan, New York University Abu Dhabi, United Arab Emirates, talal.rahwan@nyu.edu

Pooja S. B. Rao, International Institute of Information Technology Bangalore, pooja.rao@iiitb.ac.in

Claire E. Robertson, Colby College, claire.robertson@colby.edu

David M. Rothschild, Microsoft Research, David@ResearchDMR.com

Matthew J. Salganik, Princeton University, mjs3@princeton.edu

Eric Schulz, Institute for Human-Centered AI, Helmholtz Computational Health Center, Munich, Germany, eric.schulz@helmholtz-munich.de

Chirag Shah, University of Washington, chirags@uw.edu

Yash Raj Shrestha, University of Lausanne, yashraj.shrestha@unil.ch

Ekaterina Shutova, University of Amsterdam, e.shutova@uva.nl

Alexandra A. Siegel, University of Colorado Boulder, alexandra.siegel@colorado.edu

Almog Simchon, Department of Psychology, Ben-Gurion University of the Negev, Israel, almogsi@post.bgu.ac.il

Huan Sun, The Ohio State University, sun.397@osu.edu

Malte Toetzke, Technical University of Munich & Max Planck Institute for Innovation and Competition, malte.toetzke@tum.de

Jay J Van Bavel, New York University; Norwegian School of Economics, jay.vanbavel@nyu.edu

Michelle Vaccaro, MIT, Institute for Data, Systems, and Society (IDSS), vaccaro@mit.edu

Jennifer Wortman Vaughan, Microsoft Research, jenn@microsoft.com

Effy Vayena, Health Ethics and Policy Lab, ETH Zurich, effy.vayena@hest.ethz.ch

Pedro O.S. Vaz-de-Melo, Universidade Federal de Minas Gerais, Brazil, olmo@dcc.ufmg.br

Briana Vecchione, Data & Society Research Institute, briana@datasociety.net

Angelina Wang, Cornell Tech, angelina.wang@cornell.edu

Robert West, EPFL, robert.west@epfl.ch

Robb Willer, Stanford University, willer@stanford.edu

Dirk U. Wulff, Center for Adaptive Rationality, Max Planck Institute for Human Development & Faculty of Psychology, University of Basel, wulff@mpib-berlin.mpg.de

Renwen Zhang, Nanyang Technological University, Singapore, renwen.zhang@ntu.edu.sg

Simone Zhang, New York University, simone.zhang@nyu.edu

Steve Rathje, New York University, srathje@alumni.stanford.edu

Manoel Horta Ribeiro, Department of Computer Science, Princeton University, manoel@cs.princeton.edu

Correspondence: feuerriegel@lmu.de

Teaser (max. 330 characters)

Large language models (LLMs) offer new opportunities for behavioral and social science, but their rapid evolution poses challenges for research rigor. We introduce a consensus-based reporting checklist to improve transparency, reproducibility, and ethical accountability of LLM-based research in the behavioral and social sciences.

Main (max. 2000 words)

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 behavioral and social sciences are enthusiastic about the potential of LLMs to open new avenues for studying human behavior [1]. 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 analysis [2]. 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 rigor.

Here, we present a consensus-based checklist, **GUIDE-LLM,** for research involving LLMs in behavioral and social science with the aim of strengthening transparency, reproducibility, and ethical accountability (https://www.llm-checklist.com/). GUIDE-LLM stands for guidelines for the use of LLMs in behavioral and social science. The GUIDE-LLM checklist provides researchers with concrete guidance on improving transparency, reproducibility, and ethical use, strengthening rigor 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 LLMs are implemented, used, and reported in scientific research can vary widely. For example, the label “ChatGPT” can refer to different underlying models (e.g., GPT-4, GPT-4o), each with multiple versions often marked by timestamps (e.g., gpt-4o-2024-11-2). Other LLMs, such as Llama, even within the same version number, are available in different sizes (e.g., 8B vs. 70B parameters), leading to large performance differences across tasks. Even with the same model, behavior may differ depending on the access mode (e.g., via the official API or a Web interface), due to differences in system prompts (specific prompts with predefined instructions before any user input to control the model’s behavior) and runtime environments (e.g., 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, making long-term reproducibility and verification difficult [6].

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, while lower values generate more deterministic outputs. Such randomness can substantially influence behavior, especially in tasks involving text generation or reasoning. Even with a temperature set to zero, outputs can still vary due to hardware-level non-determinism (note: low temperatures are not necessarily preferable but depends on the research objective such as benefits from greater exploration vs. 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 (e.g., 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 variability [7]. Slight differences in how prompts are phrased can drastically change model outputs. One can steer an LLM's behavior 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, making 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 [8,9], including stigmatizing and stereotypical language. For downstream research using LLMs in behavioral 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 behavioral and social science may overlook issues such as the propagation of societal biases or data contamination (e.g., prior exposure of LLMs to test materials during training), thereby complicating the interpretation and robustness of findings that involve LLMs. Researchers must also account for LLMs with memory, where prior exposure to study materials in previous sessions may influence subsequent outputs and which makes reproducibility challenging.

Consensus method

The checklist was developed through a preregistered, two-round, reactive Delphi study [5], with input from an expert panel of researchers experienced in conducting LLM-based studies across various subdisciplines from behavioral and social sciences (e.g., psychology, political science, economics, sociology) as well as machine learning researchers specialized in natural language processing. In Round 1, an expert panel (*N*=68*)* evaluated an initial pool of candidate items using a five-point Likert-type scale ranging 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 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 but also additional experts to improve geographic representation), but now with a binary inclusion vote (“include” vs. “exclude”). For details, see the Supplementary Materials.

We used a preregistered consensus threshold of more than two-thirds of votes in favor of inclusion, so that all items reach broad expert support for their inclusion. Items not meeting 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.

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 (see Table 1). Overall, the GUIDE-LLM checklist comprises 14 items, covering the scope of LLM use (2 items), model/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 behavioral 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, while 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/system (e.g., 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 specific analytic tasks such as identifying sentiment, detecting emotions, or classifying moral language, so human validation helps verify that the

LLM’s responses indeed reflect the intended construct [3]. 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 where 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 resources [10], the checklist urges transparent disclosure of any potential competing interests. Clear reporting of such relationships helps readers assess possible sources of bias and ultimately strengthens trust in LLM-based behavioral science research.

#	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 vs. fully automated): ...
Model/system details	
Item B.1	Model name, including provider, model size, exact version/ID, date of access, and source link (if possible): ...
Item B.2	Model access (e.g., API, web interface, local) and context mode (e.g., chat mode or separate calls): ...
Item B.3	Relevant LLM configuration(s) reported (as applicable), such as temperature, max 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 & privacy	
Item D.1	Handling of personal or sensitive data (if any) (e.g., consent for data processing): ...
Validation & interpretation	
Item E.1	Human validation of LLM outputs: ...
Item E.2	Describe any relevant post-processing (e.g., filtering in case of format mismatches, unit conversions etc.) ...
Reproducibility	
Item F.1	Code/notebooks/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): ...

Table 1. The GUIDE-LLM checklist for promoting transparency, reproducibility, and ethical accountability in studies involving LLMs. A template for the GUIDE-LLM checklist can also be downloaded from http://www.llm-checklist.com/. The answer options are omitted from the above table and provided in the template.

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/methods (eg., whether the study conclusions are unique to a specific LLM or are generalizable across different LLMs), (4) risks of training data leakage (e.g., to reflect on cases where LLMs may have been exposed to test materials during training, potentially conflating performance estimates), (5) assessment of potential bias or systematic differences in LLM behavior that could affect the study's conclusions (e.g., LLMs are known to display gender, racial, or cultural bias that could affect results), (6) conversation transcripts, (7) a discussion of ethical implications (e.g., LLM-specific ethical considerations such as when chatbots are used for participant-facing interventions), and (8) computational resource use (e.g., to help other researchers assess reproducibility and feasibility).

Practical considerations

The checklist is intended to serve a wide range of stakeholders: behavioral 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 rigor of LLM-based research. The GUIDE-LLM checklist is not intended to prescribe specific modeling choices for LLM-based research, but rather to promote transparency.

While the majority of items in the GUIDE-LLM checklist should be broadly applicable across research designs, some flexibility is necessary. We therefore advocate a pragmatic and context-sensitive use of the checklist. Certain items may not apply to all contexts. In such cases, researchers may leave items blank or may explain why they are inapplicable. Further, the checklist is aimed at cases where LLM use is integral to the research design rather than minor or purely editorial uses of LLMs (e.g., basic language editing). 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 (e.g., source code, configuration files, or structured descriptions of the interaction logic). Overall, the checklist is intended to simplify and standardize the task for researchers by promoting transparency and reproducibility without adding unnecessary burden. To simplify reporting, authors are encouraged to reference the specific sections, pages, or appendices of their manuscripts where the required information is already provided, rather than duplicating text solely to complete the checklist.

Still, we acknowledge the limitations of the current checklist, which reflect the rapid pace at which LLM technology evolves. As such, the checklist is maintained as a "living document" on the website 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 such as vision-language models, and to AI agents that perform tasks (semi-)autonomously. Further, reporting parameters such as temperature, seeds, or other configuration settings should not be interpreted as guaranteeing full reproducibility—particularly for closed-source models—but rather as a necessary first step toward greater transparency.

Data availability

The anonymized raw data as well as the survey are publicly available via https://osf.io/mv63j. The methodology and analysis plan were preregistered before the project at: https://osf.io/9zgva

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.S., D.Y.) prepared the first version of the checklist items and reviewed the feedback from the Delphi study. S.F. conducted the survey study, analyzed 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.

Competing interests

Tim Althoff has received funding from Google, Microsoft, and Apple, and is a part-time Staff Research Scientist at Google. Christopher Barrie has received funding from OpenAI. Hal Daumé has received funding from Amazon, Google, Meta, and the Schmidt Foundation, and was previously employed at Microsoft Research. Morteza Dehghani has received funding from Google and Microsoft, and is a part-time Senior Research Scientist (Contractor) at Apple. Laura K. Globig has received funding from Google. Tom Griffiths has received funding from OpenAI, Google, and Meta. Oliver Hauser has collaborated with Google Deepmind and received funding from Anthropic, UKRI and the UK AI Security Institute. Tiancheng Hu has received funding from Apple, Google, and OpenAI. Anna Ivanova has received funding from Ferguson Control Systems (via NSF BRAIN). Himabindu Lakkaraju 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. Steve Rathje and Jay Van Bavel have received research funding from OpenAI and Google. Ekaterina Shutova has received funding from Google, Meta, and Deloitte. Huan Sun has received funding from Amazon, Google, Cisco, OpenAI, Fujitsu, and Intuit. Rada Mihalcea has received funding from OpenAI and LG AI, and in-kind support from Amazon and Microsoft. Almog Simchon has received funding from Google. Manoel Horta Ribeiro has received funding from Meta and OpenAI. Alice Oh received funding from Google, Microsoft, LG AI, and Naver. Alice Oh is a Faculty Visiting Researcher at Google, but this work was done in her capacity as a faculty member at KAIST. The other authors declare no competing interests.

Funding

Tim Althoff was supported in part by NSF IIS-1901386, NSF CAREER IIS-2142794, and NIH R01MH125179. Oliver Hauser is grateful to the UKRI Future Leaders Fellowship (project reference: 1067). Himabindu Lakkaraju was supported in part by NSF Awards IIS-2008461, IIS-2040989, and IIS-2238714; and an AI2050 Early Career Fellowship from Schmidt

Sciences. Rada Mihalcea was supported in part by NSF IIS-2306372. Tanmoy Chakraborty was supported by the Rajiv Khemani Young Faculty Chair Professorship in AI. M.J. Crockett was supported by a grant from the Princeton Catalysis Initiative.

References (max. 10)

[1] Messeri, L., & Crockett, M. J. Artificial intelligence and illusions of understanding in scientific research. *Nature* **627**, 49–58 (2024).

[2] Feuerriegel, S., et al. Using natural language processing to analyse text data in behavioural science. *Nat Rev Psychol* **4,** 96–111 (2025).

[3] Argyle, L. P., et al. Leveraging AI for democratic discourse: Chat interventions can improve online political conversations at scale. *Proceedings of the National Academy of Sciences* **120**, e2311627120 (2023).

[4] Costello, T. H., Pennycook, G., & Rand, D. G. Durably reducing conspiracy beliefs through dialogues with AI. *Science* **385**, eadq1814 (2024).

[5] Grime, M. M. & Wright, G. Delphi Method. Wiley Statistics Reference Online, 1–6 (2016).

[6] Palmer, A., Smith, N.A., & Spirling, A. Using proprietary language models in academic research requires explicit justification. In: *Nat Comp Sci* **4**, 2–3 (2024).

[7] Yang, C., Wang, X., Lu, Y., Liu, H., Le, Q. V., Zhou, D., & Chen, X. Large language models as optimizers. In: *International Conference on Learning Representations* (2023).

[8] Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. On the dangers of stochastic parrots: Can language models be too big? In: *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency*, pp. 610–623 (2021).

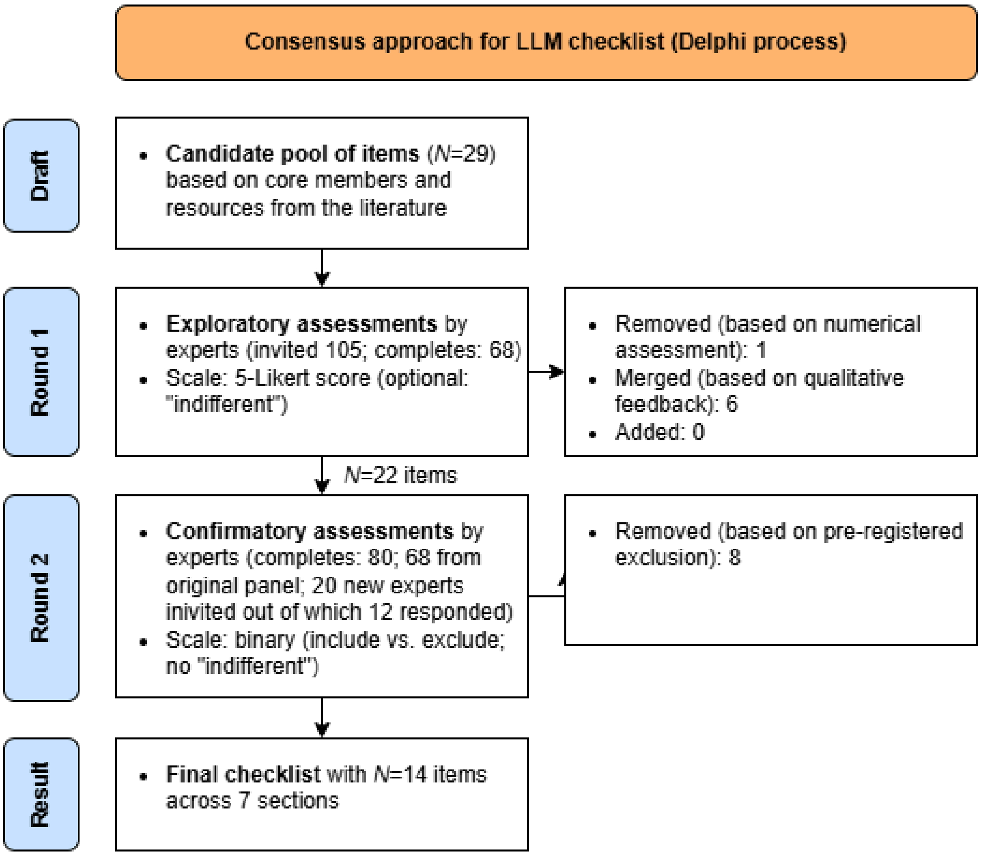
[9] Hu, T., Kyrychenko, Y., Rathje, S., Collier, N., van der Linden, S., & Roozenbeek, J. Generative language models exhibit social identity biases. *Nat Comp Sci* **5**, 65–75 (2025).

[10] Ahmed, N., Wahed, M., & Thompson, N. C. The growing influence of industry in AI research. *Science* **379**, 884–886 (2023).

Supplementary Materials

Overview

To develop the GUIDE-LLM reporting checklist for studies using large language models (LLMs) in behavioral and social science, we employed a two-round reactive Delphi process (Supplementary Fig. 1). The Delphi process is widely used as a research methodology to establish consensus among experts on best practices and reporting standards in scientific research. For details on the Delphi methodology, see Grime & Wright (2016), Taylor (2020), and McKenna (1994). In our study, the Delphi process consisted of two rounds designed to identify, refine, and reach consensus on checklist items. In the first round, an expert panel of researchers with expertise spanning behavioral and social sciences as well as machine learning and natural language processing evaluated a preliminary pool of candidate items derived from prior literature and existing reporting frameworks. Based on their quantitative ratings and qualitative feedback, the checklist was iteratively revised by adding, merging, removing, or rewording items to improve clarity and relevance. In the second round, the revised items were again assessed by an expert panel to vote on the final inclusion. Items that achieved consensus support were retained to form the final version of the checklist.


Consensus approach for LLM checklist (Delphi process)
Draft
Round 1
Round 2
Result
• Candidate pool of items (N=29) based on core members and resources from the literature
• Exploratory assessments by experts (invited 105; completes: 68)
• Scale: 5-Likert score (optional: "indifferent")
• Removed (based on numerical assessment): 1
• Merged (based on qualitative feedback): 6
• Added: 0
N=22 items
• Confirmatory assessments by experts (completes: 80; 68 from original panel; 20 new experts inivited out of which 12 responded)
• Scale: binary (include vs. exclude; no "indifferent")
• Removed (based on pre-registered exclusion): 8
• Final checklist with N=14 items across 7 sections

Supplementary Figure 1. Study flow diagram.

Our checklist should be viewed as complementary to existing documentation efforts in machine learning, such as Datasheets for Datasets (Gebru et al., 2021), Model Cards

(Mitchell et al., 2019), submission checklists for some ML/NLP conferences (e.g., Dodge et al., 2019), and emerging model sheets released for some contemporary frontier LLMs. These initiatives focus primarily on documenting the model itself (i.e., the data sources, capabilities, limitations, and evaluation). By contrast, GUIDE-LLM addresses the researcher's use of an LLM within a behavioral or social science study by emphasizing transparency in study design, prompts, configurations, validation, and interpretation. While improved model-level documentation is an important step toward reproducibility and responsible development of LLMs, the idea of GUIDE-LLM is to ensure rigor in downstream scientific applications.

In developing GUIDE-LLM, we also reflected on critiques in the meta-literature (e.g., Magnusson et al., 2023, Thomassen et al., 2014) suggesting that reporting checklists can become overly complex, burdensome, or encourage superficial box-ticking rather than meaningful transparency. This concern was also echoed in the qualitative feedback we received throughout the Delphi process, where experts repeatedly stressed the importance of keeping the checklist concise and minimizing unnecessary effort. Eventually, this was also reflected in the Delphi process that narrowed the initial 29 candidate items to 14 items. A key distinction of GUIDE-LLM relative to other, developer-oriented checklists (e.g., Dodge et al., 2019; Mitchell et al., 2019) is that GUIDE-LLM is tailored for domain researchers rather than method developers, who may otherwise overlook or be unaware of key technical details when reporting their methods. Although any checklist can, in principle, be gamed or reduced to rote box-ticking, our experience—such as supervising students and reviewing manuscripts—shows that even essential details (e.g., exact model versions or prompts) are frequently omitted in practice. GUIDE-LLM therefore aims to provide a minimal but effective set of reminders that help prevent common oversights while remaining feasible and useful across behavioral and social science applications.

Differences from other reporting checklists

GUIDE-LLM differs from existing reporting frameworks. To illustrate the differences, we compare it with three widely used checklists: (i) TRIPOD-LLM (Gallifant et al., 2025), (ii) Model Cards (Mitchell et al., 2019), and (iii) REFORMS (Kapoor et al., 2024).

TRIPOD-LLM (Gallifant et al., 2025) extends prior TRIPOD checklists to address studies involving LLMs in healthcare. The primary focus is to emphasize transparent reporting across medical settings, particularly regarding the deployment context of LLMs in healthcare, model versioning, fine-tuning procedures, evaluation metrics, human oversight, bias assessment, and clinical task performance. As such, many TRIPOD-LLM reporting items are tailored to biomedical applications and may not directly translate to LLM-based studies in behavioral and social science. For example, TRIPOD-LLM focuses on reporting the exact medical tasks, characteristics of the patient population, the intended use in the care pathway, the involvement of patients and the public during the design, and the evaluation approach, including usability in clinical environments. In contrast, many LLM applications in behavioral and social science are not covered by TRIPOD-LLM, such as using LLMs to generate experimental stimuli, simulate human judgments in silico, annotate text for latent constructs (e.g., emotions, moral language), model social interactions, or deliver conversational interventions. In these contexts, methodological transparency hinges more critically on detailed documentation of prompt design, system instructions, sampling parameters (e.g., temperature), and post-processing decisions, which thus motivate a tailored checklist.

Model Cards (Mitchell et al., 2019) are designed to describe the development and properties of a machine learning model itself, including its training data, intended use, benchmarks, and known limitations or biases. GUIDE-LLM is conceptually different from Model Cards in that it does not focus on documenting the model as a technical artifact, but rather on documenting how researchers use a model within a specific empirical study. For example, a Model Card may report that a model was trained on web-scale text and evaluated on machine learning benchmarks. However, many LLM applications in behavioral and social science do not involve model development; instead, researchers access models via APIs or web interfaces, which introduces different reporting needs. Moreover, Model Cards are designed for general machine learning systems and are not tailored to LLM-specific practices such as prompting, system instructions, chain-of-thought elicitation, memory properties, or agentic frameworks. As a result, Model Cards do not capture many aspects that are central to behavioral and social science applications, including whether (hidden) system prompts were active, the exact prompt formulations used, parameter settings (e.g., temperature, maximum tokens, number of runs), or study-specific post-processing decisions. GUIDE-LLM therefore operates at the study level rather than the model level, thus encouraging transparency and reproducibility about how LLMs are actually implemented within empirical research designs.

REFORMS (Kapoor et al., 2024) is designed for machine-learning–based science more broadly and does not include LLM-specific elements nor dimensions tailored to behavioral and social science applications. In contrast, our GUIDE-LLM checklist serves a different purpose by focusing explicitly on LLM-related reporting aspects (such as prompt documentation, context mode, temperature, etc.) that are central to LLM-based research but not addressed in general ML checklists. Further, we focus explicitly on the tasks and thus reporting needs common in behavioral and social science.

Taken together, the focus on behavioral and social science reflects unique methodological and theoretical features of these disciplines, and implies that LLMs are often embedded directly into the research design rather than serving solely as predictive tools. LLMs may, for example, generate persuasive messages, moral dilemmas, or social scenarios; act as simulated participants in in silico experiments; annotate latent psychological constructs; or interact with human participants as interventions. In these settings, small variations in prompt phrasing, system instructions, temperature settings, or filtering rules can influence theoretical conclusions about cognition, judgment, or social behavior. Moreover, behavioral research frequently relies on human validation to assess construct validity (e.g., by evaluating whether LLM outputs meaningfully capture constructs such as emotion, morality, or identity; see Feuerriegel et al., 2025), which introduces reporting needs that differ from those in purely technical or clinical domains.

Procedure

Before the start of the project, the research plan was preregistered on September 19, 2025 via https://osf.io/9zgva. Further, ethical approval was obtained from LMU Munich School of

Management (ETH-SOM-022). A CREDES reporting checklist[1] is available as a supplementary file.

Preparation

Before the start of the Delphi process, the core team (Christopher Barrie, M.J. Crockett, Stefan Feureriegel, Laura K. Globig, Killian L. McLoughlin, Dan-Mircea Mirea, Arthur Spirling, Diyi Yang, Steve Rathje, Manoel Horta Ribeiro) developed an initial pool of candidate items for a reporting checklist aimed at LLM use in behavioral and social sciences. The aim at this stage was to create a broad and comprehensive item pool that would cover the full range of potential reporting dimensions, thereby reducing the need for substantial additions in later rounds. This preliminary draft was informed by the core team's own experience as author, reviewer, and editor, but also three main sources: (i) prior behavioral and social science studies that employed large language models as research tools, (ii) methodological and "how-to" papers outlining best practices for using LLMs in behavioral science (for example, Demszky et al., 2023; Feuerriegel et al., 2025), and (iii) established reporting guidelines and checklists from adjacent domains such as MI-CLAIM, TRIPOD-LLM, and the 2025 version of the NeurIPS Paper Checklist. As in other reporting checklists, the items were intentionally phrased in a concise and accessible manner to facilitate use and reduce burden among end-users. At the same time, each item was accompanied by a brief explanatory note clarifying terminology or providing illustrative context. The draft checklist was iteratively revised and refined within the core team until consensus was reached among all members. This draft then served as the input to Round 1 of the Delphi process.

Overall, the draft had 29 items and is available via https://osf.io/mv63j/. The draft (and thus the survey) was grouped into thematic areas to ensure comprehensive coverage but also easier navigation across relevant dimensions. These areas correspond to the major sections of the checklist: Scope of LLM Use (3 items); Model/System Details (5 items); Prompts & Parameters (5 items); Data Inputs & Privacy (3 items); Validation & Interpretation (4 items); Bias, Fairness & Safety Evaluation (2 items); Reproducibility (2 items); Ethics & Governance (3 items); and Compute Cost (2 items).

Round 1

For the first round of the Delphi process (September 19–October 26, 2025), the core team identified and invited domain experts through professional networks and targeted searches of the relevant literature in behavioral and social sciences, as well as machine learning, particularly with expertise focused on the application and evaluation of LLMs as well as AI transparency. In doing so, we screened recent publications in general outlets (e.g., *Nature*, *Nature Human Behaviour, PNAS*), which frequently publish research leveraging technological innovations for behavioral and social science, as well as selected domain-specific journals to identify researchers actively publishing LLM-based work. Eligible participants were researchers with documented experience in the methodological, ethical, or applied use of LLMs—for example, first or last authors of peer-reviewed publications or

[1] https://www.equator-network.org/reporting-guidelines/credes/

preprints demonstrating credible and technically rigorous applications or evaluations of LLMs in behavioral or social science contexts. Invited experts were also given the opportunity to suggest additional qualified participants; however, the vast majority of suggested names had already been invited.

The panel was intentionally diverse, combining expertise from junior as well as senior scholars (to recognize that junior scholars often serve as lead authors on computational works and often bring substantial hands-on expertise), as well as from different subfields such as psychology, neuroscience, political science, sociology, communication science, and management, together with methodological expertise from computer science and natural language processing. To ensure sufficient representation of methodological experts, we preregistered a minimum of >5 experts with a primary background in ML/NLP. The survey is available via https://osf.io/mv63j/

Invitations were sent via personalized email. Participation was voluntary, but participants were offered co-authorship if they completed both rounds of the Delphi process. The survey was designed via Google Forms. Our preregistered minimum target was 20 complete responses, which was informed by other checklists (e.g., TRIPOD-LLM) but with no upper cap on the number of contributors. Overall, we invited 105 experts (including the core members), out of whom 68 successfully completed the survey in Round 1.

Overall, according to the Frascati Manual classification (OECD, 2015), 36.8% of respondents identified as (A) *top-grade researchers* (e.g., full professors or equivalent senior positions), 17.6% as (B) *senior researchers* (e.g., associate professors), 30.9% as (C) *recognized researchers* (e.g., assistant professors, post-docs), and 14.7% as *first-stage researchers* (e.g., doctoral candidates or early-career researchers). Experts reported having diverse disciplinary backgrounds (multiple answers were allowed). Frequent field(s) of expertise were computational social science (55.9%), AI/ML (54.4%), natural language processing (44.1%), psychology (33.8%), human-computer interactions (32.4%), cognitive science (19.1%), political science (16.2%), ethics or governance (14.7%), economics (10.3%), public policy (10.3%), management (10.3%), communication (7.4%), sociology (5.9%), and neuroscience (2.9%).

Experts who agreed to participate received an online survey containing the full list of preliminary checklist items and accompanying explanatory notes. The survey began with an overview of the project's goals and key design considerations—namely, that the checklist aimed to capture a minimum set of core items essential for transparent, reproducible, and ethically accountable research using LLMs in behavioral and social science. Participants were reminded that the Delphi process was consensus-oriented, not prescriptive, and that the final checklist would serve as a flexible framework that different subfields could adapt to their specific needs. Each item was rated on a five-point Likert-type scale ranging from "strongly exclude" (–2) to "strongly include" (+2). The choice for the Likert-type scale is based on other checklists (e.g., CONSORT, TRIPOD AI, Aczel et al., 2020). Participants could also select "indifferent" if they felt unsure or lacked relevant expertise. The instructions emphasized that the task was not about rating all items equally but rather about identifying relative priorities, recognizing that not all items are relevant across all research designs. Thus, participants were encouraged to reflect carefully on which items are most pertinent for the field. Below each item, participants could provide open-ended feedback such as suggestions for rewording, alternative formulations, or entirely new items, along with optional justifications. At the end of the survey, an additional open-ended text box allowed participants to propose any further items or considerations not covered in the main list.

The primary metric for Round 1 was the Relevance Score (RS) for each item, defined as the mean rating across respondents excluding “indifferent” responses (see Supplementary Table 1). Items with a positive RS were provisionally retained, while all qualitative feedback was reviewed in detail to identify suggestions for merging, clarifying terminology, or addressing overlooked aspects. The core members reviewed all feedback in anonymized form. Following an initial analysis, the core members convened on October 27, 2025 followed by discussions via email to incorporate the quantitative and qualitative feedback and to deliberate on item-specific comments, redundancy across sections, and opportunities for clarity. Based on the quantitative ratings and qualitative feedback from Round 1, several changes were made as follows:

- **Deletions:** One item (for reporting the environmental impact of LLMs; RS = –0.5) was deleted because of a low Relevance Score (RS).
- **Merging:** Additional items were integrated into other items due to the qualitative feedback: one item asking where LLMs were used in the research process was perceived as duplicative of Item A.1; an item on data retention was integrated into the broader section on handling of personal or sensitive data (Item D.1); an item on safety layers and moderation systems was merged into the more general category on LLM customization (Item B.4): an item around risk evaluation and discussing safeguards against harmful or adverse outputs was merged into a general item on ethical considerations (Item H.1); an item asking about types of data provided to the LLM including sensitive data was removed given that a separate item already focuses on the step of handling personal or sensitive data (Item D.2); and one item concerning participant disclosure and debriefing was removed, as it was considered a general research ethics issue rather than one specific to LLM-based studies (this aspect, however, was incorporated into the broader ethical dimension under Item H.1 to ensure that participant-facing ethical considerations remained represented).
- **Revisions:** One separate item asking about IRB/ethics approval was revised to focus on the broader ethical implications of the research (Item H.1), since ethics approval was regarded by the experts as standard scientific practice rather than specific to LLM-based studies. The item on the justification for model(s) choice (Item B.6) was expanded to incorporate additional dimensions identified by experts, such as cost and ease of use. Likewise, the item on potential risk of bias was rephrased to focus on potential risk of bias or other systematic differences in LLM behavior that could affect the study’s conclusions (Item F.1) to better capture diverse sources of bias beyond demographic characteristics. In five cases, the response format was modified (for example, replacing fixed tick boxes with free-text fields) to allow end-users to provide more nuanced input.
- **Restructuring:** Some items were moved to different areas; specifically, Sections B and C were reorganized so that the former focuses on the model/system, while the latter is exclusively reserved for prompt-related aspects.
- **Editorial edits:** Finally, minor editorial and orthographic revisions were introduced to ensure consistency in terminology and scope. For example, phrasing was standardized to refer jointly to both behavioral and social science, the notation

“LLM(s)” and “model(s)” was used to clarify that multiple systems could be used and reported within a single study, and expressions such as “if any” were added where applicable to indicate that certain items (for example, those concerning sensitive data) may not apply to all studies.

In Round 1, only few suggestions for additional items were made, which were ultimately discarded after careful discussion, but which can be attributed to the already large set of candidate items and the extensive merging and reformulation undertaken after Round 1. As a result, the revised checklist had 22 items that served as input to Round 2.

Supplementary Table 1. Results from Round 1. Item names are abbreviated for better readability. Votes for Likert scales and “Indifferent” reported as percentages.

#	Item	Relevance score (RS)	Strongly include (+2)	(+1)	Neutral (0)	(-1)	Strongly exclude (-2)	Indifferent
Item A.1	LLMs were used in this project for: ...	1.7	73.1	25.4	1.5	0.0	0.0	1.5
Item A.2	Research stage(s) where LLMs were used: ...	0.7	33.8	30.8	13.8	15.4	6.2	4.4
Item A.3	Degree of automation (human-in-the-loop vs. fully automated): ...	0.9	40.3	32.8	13.4	4.5	9.0	1.5
Item B.1	Model name, including provider, model size, exact version/ID, date of access, and source link (if possible): ...	1.8	88.2	8.8	1.5	1.5	0.0	0.0
Item B.2	Model access (e.g., API, web interface, local) and context mode (e.g., chat mode or separate calls): ...	1.3	53.0	31.8	7.6	6.1	1.5	2.9
Item B.3	Customization: ...	1.3	53.7	29.9	10.4	3.0	3.0	1.5
Item B.4	Justification for the model choice along the following dimensions: ...	0.7	30.8	33.8	18.5	12.3	4.6	4.4
Item B.5	Additional safety layers/moderation systems (e.g., provider filters, custom guards): ...	0.5	18.6	33.9	27.1	15.3	5.1	13.2
Item C.1	System-wide instructions (if any): ...	1.2	48.5	33.3	12.1	4.5	1.5	2.9
Item C.2	Did the LLM session include persistent memory across interactions? ...	1.1	41.2	39.7	13.2	2.9	2.9	0.0
Item C.3	Exact prompt(s) reported: ...	1.7	80.9	16.2	0.0	1.5	1.5	0.0
Item C.4	Discussion of the rationale for the prompt design: ...	0.4	19.4	28.4	25.4	22.4	4.5	1.5
Item C.5	Relevant LLM configuration reported (as applicable), such as temperature, max tokens, seed, and number of runs: ...	1.2	47.8	37.3	9.0	1.5	4.5	1.5
Item D.1	Categories of data provided to the LLM described (e.g., raw text, transcripts, personally identifiable information [PII]): ...	1.0	35.8	35.8	20.9	3.0	4.5	1.5
Item D.2	Handling of personal or sensitive data (e.g., consent for data processing): ...	1.0	49.3	22.4	10.4	10.4	7.5	1.5
Item D.3	Data retention and data usage by vendor (e.g., use of data for training, human review, or long-term storage): ...	0.5	22.4	37.3	20.9	10.4	9.0	1.5
Item E.1	Human validation of LLM outputs: ...	1.3	62.7	20.9	7.5	6.0	3.0	1.5
Item E.2	Comparison against other methods/LLMs: ...	0.4	20.6	29.4	22.1	20.6	7.4	0.0
Item E.3	Describe any relevant post-processing (e.g., filtering in case of format mismatches, unit conversions etc.): ...	0.7	28.4	34.3	19.4	13.4	4.5	1.5
Item E.4	Training data leakage risks addressed: ...	0.7	26.2	41.5	13.8	9.2	9.2	4.4
Item F.1	Assessment of bias, stereotypes, or other forms of systematic discrimination in LLM outputs: ...	0.3	25.8	24.2	21.2	13.6	15.2	2.9
Item F.2	Safeguards for harmful/adverse outputs and risk evaluation: ...	0.4	24.2	25.8	24.2	16.7	9.1	2.9
Item G.1	Code/notebooks/scripts for LLM calls shared: ...	1.3	54.5	28.8	7.6	9.1	0.0	2.9
Item G.2	Conversation transcripts: ...	0.6	25.4	26.9	32.8	11.9	3.0	1.5
Item H.1	IRB/ethics review: ...	0.9	54.0	9.5	19.0	9.5	7.9	7.4
Item H.2	Funding, support, or other relevant relationships (including in-kind access to compute or models, or professional affiliations): ...	1.0	49.2	23.1	9.2	15.4	3.1	4.4
Item H.3	Participant disclosure and debriefing: ...	1.1	50.0	27.3	13.6	4.5	4.5	2.9
Item I.1	Computational resources (e.g., API call counts, tokens, financial costs, or compute time): ...	0.3	16.4	34.3	22.4	14.9	11.9	1.5
Item I.2	Environmental impact: ...	-0.5	9.4	18.8	21.9	17.2	32.8	5.9

Round 2

For the second round (November 5–March 11, 2026), all experts who had participated in Round 1, including the core team members, were again invited via email. In addition, to improve the representativeness and disciplinary diversity of the panel, we recruited additional experts using the same eligibility criteria as in Round 1. These individuals were identified through a snowball approach based on recommendations from Round 1 participants. As a result, and deviating from the original preregistration, 12 additional experts participated in Round 2 (invited: 20 experts; response rate: 60%).[2] The survey was again implemented via Google Forms. The survey is available via https://osf.io/mv63j/. Overall, we received 80 responses.

Owing to the expanded expanded panel in Round 2, the composition of the participants was as follows. Overall, 36.2% of respondents identified as (A) *top-grade researchers* (e.g., full professors or equivalent senior positions), 22.5% as (B) *senior researchers* (e.g., associate professors), 28.8% as (C) *recognized researchers* (e.g., assistant professors, post-docs), and 12.5% as *first-stage researchers* (e.g., doctoral candidates or early-career researchers). Experts reported having diverse disciplinary backgrounds (multiple answers were allowed). Frequent field(s) of expertise were computational social science (66.2%), AI/ML (61.8%), natural language processing (51.5%), human-computer interactions (36.8%), and psychology (35.3%).

For this round, the revised version of the checklist was presented, containing all items that had been retained or reformulated after Round 1. Each item was again accompanied by a short note providing additional clarifications or justifications. The survey instructions were updated to clarify that Round 2 focused on a binary decision; that is, experts were asked to indicate whether each item should be "included" or "excluded" from the final checklist. Following the preregistration, there was no field for abstentions or "indifferent" responses, as the goal was to reach consensus on a set of items relevant across all behavioral and social science subfields. Here, we preregistered that all items will be retained where "p_include" (i.e., the proportion of "include" votes, i.e., p_include = n_include / (n_include + n_exclude) is larger than 2/3, which is similar to other checklists such as Aczel et al., 2020). The introductory section of the survey explicitly stated that the second Round 2 had a strict, predefined inclusion threshold (specifically, the rate of "include" votes must be larger than 2/3 of all votes) for an item to be retained.

The results for the second round are reported in Supplementary Table 2. As a result, 8 items that did not meet the inclusion threshold were deleted. These were: Justification for the model(s) choice (Item B.6); Discussion for the rationale of the prompt (Item C.3); Comparison against other methods/LLMs (Item E.3); Training data leakage risks addressed (Item E.4); Potential risk of bias or other systematic differences in LLM behavior that could affect the study's conclusion (Item F.1) Conversation transcripts (Item G.2); Discussion of

[2] By design, Round 2 focused on binary inclusion decisions and therefore did not formally provide newly recruited experts with the opportunity to propose additional items or suggest changes to the wording wording. However, all newly recruited experts were asked to make such suggestions as separate comments; however, no such suggestions for additional items or major wording revisions were submitted. This reflects our observations from Round 1, where also no new items were suggested, implying that the selected items are already sufficiently comprehensive and that their wording is generally acceptable to the expert panel.

relevant ethical implications of research (formerly Item H.2); and Computational resources (formerly Item I.1). These idems did not meet the inclusion threshold. The remaining 14 items achieved consensus and were retained in the final version of the checklist (see main paper). The category “Ethics” was renamed to “Competing interest” to better reflect the included items. Note that, after exclusion, the items were re-numbered, so the numbering in the above is different from the numbering in the main paper.

In comparing the outcomes of Rounds 1 and 2, we observed that several items that had received high relevance ratings in the first round did not reach the inclusion threshold in the second round. Although we did not formally collect data on the reasons for these changes, informal communications with several experts indicated that their decisions in Round 2 were not driven by a perceived lack of importance or relevance of these items. Instead, experts emphasized that their votes reflected a stronger focus on generalizability—that is, prioritizing items that are applicable across a broad range of LLM-based studies in the behavioral and social sciences. Consequently, they voted to exclude certain items that they viewed as valuable but context-specific or that would be covered by other mechanisms (e.g., IRB approval) to ensure that the final checklist captures only a minimum set of reporting items that are broadly applicable.

To explore potential heterogeneity in expert agreement within specific subdisciplines (e.g., psychology, AI/ML, computational social science), we conducted additional comparisons and analyzed the discipline-specific assessments. The analyses indicated generally consistent agreement within subfields. Overall, no systematic disciplinary differences in agreement were observed that were particularly noteworthy. Notwithstanding, this pattern may simply reflect that the final set of minimum reporting items is broad and widely applicable across disciplines. It does not, however, preclude the possibility that discipline-specific extensions or adaptations of the checklist may be warranted in the future (e.g., with dedicated sections for ethical considerations when LLMs are used to guide participant-facing interventions). We also qualitatively inspected the responses of newly recruited Round 2 participants compared to those who had participated in Round 1 but did not observe any notable differences in voting patterns or overall assessments.

Supplementary Table 2. Results from Round 2. Item names are abbreviated for better readability. The ratio of “include” votes (p_include) is reported as %. The final outcome reports as to whether consensus for inclusion was achieved.

#	Name	Inclusion (%)	Final outcome
Item A.1	LLM(s) were used in this project for: ...	97.5	✅
Item A.2	Degree of automation (human-in-the-loop vs. fully automated): ...	73.8	✅
Item B.1	Model name, including provider, model size, exact version/ID, date of access, and source link (if possible): ...	100.0	✅
Item B.2	Model access (e.g., API, web interface, local) and context mode (e.g., chat mode or separate calls): ...	85.0	✅
Item B.3	Relevant LLM configuration(s) reported (as applicable), such as temperature, max tokens, seed, and number of runs. ...	93.8	✅
Item B.4	Customization: ...	83.8	✅
Item B.5	Did the LLM session(s) include persistent memory across interactions? ...	78.8	✅
Item B.6	Justification for the model(s) choice along the following dimensions: ...	53.8	❌
Item C.1	Exact prompt(s) reported: ...	98.8	✅
Item C.2	System-wide instructions (if any): ...	87.5	✅
Item C.3	Discussion of the rationale for the prompt design: ...	36.2	❌
Item D.1	Handling of personal or sensitive data (if any) (e.g., consent for data processing): ...	75.0	✅
Item E.1	Human validation of LLM outputs: ...	85.0	✅
Item E.2	Describe any relevant post-processing (e.g., filtering in case of format mismatches, unit conversions etc.) ...	77.5	✅
Item E.3	Comparison against other methods/LLMs ...	48.8	❌
Item E.4	Training data leakage risks addressed ...	56.2	❌
Item F.1	Potential risk of bias or other systematic differences in LLM behavior that could affect the study’s conclusions? ...	60.0	❌
Item G.1	Code/notebooks/scripts for LLM calls shared: ...	85.0	✅
Item G.2	Conversation transcripts: ...	66.2	❌
Item H.1	Funding, support, or other relevant relationships (including in-kind access to compute or models, or professional affiliations): ...	72.5	✅
Item H.2	Discuss relevant ethical implications of the research: ...	56.2	❌
Item I.1	Computational resources (e.g., API call counts, tokens, financial costs, or compute time): ...	55.0	❌

Finalization

To facilitate the practical use of the GUIDE-LLM checklist, we created a dedicated website that allows researchers to easily complete, view, and export the checklist online (see http://llm-checklist.com/). The website was designed to make the checklist intuitive to fill out and adaptable for integration into standard research workflows. Each item includes optional explanatory text that can be expanded for further guidance, and users can download the completed form for use in manuscript preparation or peer review.

Further, we prepared a set of explanatory notes to accompany the checklist. As in other checklists, these were not part of the Delphi process but were still developed and reviewed by the experts. We began with the explanatory material already included in the Round 1 draft and then iteratively expanded it using feedback collected during both Delphi rounds. Afterward, the Notes were revised to make them more explanatory and user-oriented, reducing technical jargon and shifting from an expert-focused style toward one accessible to a broader behavioral and social science audience. The final version of the Notes and checklist underwent an additional round of review among the expert panel to ensure clarity, completeness, and consistency with the consensus outcomes.

For the final checklist, we made some further minor stylistic changes for consistency. Specifically, we removed the answer options in Item C.1, analogous to other answer options, and replaced them with a free-text field, so that researchers can either directly copy the prompt or add the location in the paper or supporting materials. Further, we added checkboxes in the optional item on the model choice to have a consistent appearance, and we streamlined the text to use the plural forms of LLMs to avoid inconsistencies between “LLMs” versus “LLM(s)”.

During the finalization phase, we revisited the discrepancies between Rounds 1 and 2. Following careful discussion among both the core team and expert contributors, we concluded that the items excluded in Round 2—while not meeting the predefined consensus threshold—had nonetheless been rated as relevant in Round 1 and offered meaningful guidance for certain study contexts. To acknowledge their value in specific contexts while maintaining the focus on a concise set of minimal reporting standards, these items were incorporated into the online version of the GUIDE-LLM checklist as “optional” items. Researchers can therefore include these additional items when relevant to their specific study design or domain if deemed relevant.

To illustrate the intended application, we also developed a worked example based on a published behavioral science paper, demonstrating how the checklist can be used to document an LLM-based study transparently and systematically.

Future updates

To enable future updates to the GUIDE-LLM, revisions will follow a structured process. Specifically, proposed updates will be reviewed collectively by the core team at fixed intervals (e.g., annually), with major revisions (e.g., addition, removal, or redefinition of checklist items) decided through a formal vote among the core members. For substantial conceptual changes, we plan to convene a second Delphi process or equivalent consensus procedure involving the broader expert panel. This review process is designed to maintain methodological rigor and prevent arbitrary or unilateral updates, ensuring that the checklist evolves alongside advances in LLM technology and research practice.

Reflections on methodological boundaries

GUIDE-LLM is intended to surface key methodological challenges in LLM-based research and promote transparency in how they are addressed, rather than to mandate specific analytic choices or prescribe uniform standards. The checklist identifies areas where researcher discretion meaningfully shapes results, while leaving it to authors to justify and document their decisions in light of their study design and disciplinary norms. The item on human validation illustrates this approach: we do not require a particular validation protocol or threshold, but instead encourage researchers to clarify whether validation was appropriate, how it was implemented, and what its limitations are. In some contexts—such as AI-assisted literature search or large-scale exploratory workflows—full manual replication may be impractical, and human "gold standards" may themselves introduce bias. We therefore frame validation as a context-dependent consideration that should be transparently reported, thereby acknowledging that shared standards in this area are still evolving. As such, we echo the challenges inherent in human validation (e.g., large-scale human validation is resource-intensive, yet fully automated evaluation may lack construct grounding). This tension exists in research involving both human- and machine-generated data and reflects broader methodological trade-offs between scalability, reliability, and construct validity.

Rather than prescribing specific methodological standards, we acknowledge that best practices for the responsible and rigorous use of LLMs in research are still evolving. We therefore provide pointers to relevant literature outlining emerging guidance on specific dimensions of validation and LLM use more broadly (e.g., Atreja et al., 2024; Barrie et al., 2024; Feuerriegel et al., 2025; Song et al., 2020).

Ultimately, we encourage a holistic, open-science approach to transparency, reproducibility, and ethical research that extends beyond this checklist. GUIDE-LLM should further be complemented with field-specific guidelines to ensure that best practices are consistently followed across different domains of LLM use. The flexibility of LLM workflows further amplifies the risk that both data generation and analysis could be gamed through careful prompting and parameter tuning; addressing this challenge will require greater transparency and may need more formalized standards for documentation, including the use of preregistrations.

Data availability

The surveys and the anonymized responses are available via https://osf.io/mv63j/

References

Aczel, B., et al. (2020). A consensus-based transparency checklist. *Nature Human Behaviour* **4**, 4–6. Preregistration: https://osf.io/3tkhn

Atreja, S., Ashkinaze, J., Li, L., Mendelsohn, J., & Hemphill, L. (2024). Prompt design matters for computational social science tasks but in unpredictable ways. arXiv:2406.11980.

Barrie, C., Palmer, A., & Spirling, A. Replication for language models: Problems, principles, and best practice for political science. Working Paper (2024). https://arthurspirling.org/documents/BarriePalmerSpirling_TrustMeBro.pdf

Demszky, D., Yang, D., Yeager, D.S., et al. (2023). Using large language models in psychology. *Nature Reviews Psychology* **2**, 688–701.

Dodge, J., Gururangan, S., Card, D., Schwartz, R., & Smith, N. A. (2019). Show your work: Improved reporting of experimental results. In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing and the International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 2185-2194.

Feuerriegel, S., et al. (2025). Using natural language processing to analyse text data in behavioural science. *Nature Reviews Psychology* **4,** 96–111.

Gallifant, J., et al (2025). The TRIPOD-LLM reporting guideline for studies using large language models. *Nature Medicine* **31**, 60–69.

Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Iii, H. D., & Crawford, K. (2021). Datasheets for datasets. *Communications of the ACM* **64**, 86–92.

Grime, M. M. & Wright, G. (2016). Delphi Method. *Wiley Statistics Reference Online*, 1–6.

Kapoor, S., et al. (2024). REFORMS: Consensus-based Recommendations for Machine-learning-based Science. *Science Advances*, 10(18), eadk3452.

Magnusson, I., Smith, N. A., & Dodge, J. (2023). Reproducibility in NLP: What have we learned from the checklist?. In: *Findings of the Association for Computational Linguistics*, pp. 12789-12811.

McKenna, H. P. (1994). The Delphi technique: A worthwhile research approach for nursing? *Journal of Advanced Nursing* **19**:1221–1225.

Mitchell, M., et al (2019). Model cards for model reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pp. 220-229.

OECD (2015). *Frascati Manual 2015: Guidelines for collecting and reporting data on research and experimental development* (The Measurement of Scientific, Technological and Innovation Activities). Organisation for Economic Co-operation and Development, OECD Publishing. https://doi.org/10.1787/9789264239012-en

Song, H., et al. (2020). In validations we trust? The impact of imperfect human annotations as a gold standard on the quality of validation of automated content analysis. *Political Communication*, 37(4), 550-572

Taylor, E. (2020). We agree, don't we? The Delphi method for health environments research. *HERD: Health Environments Research & Design Journal* **13**, 11–23.

Thomassen, Ø., Storesund, A., Søfteland, E., & Brattebø, G. (2014). The effects of safety checklists in medicine: A systematic review. *Acta Anaesthesiologica Scandinavica* **58**, 5-18.