Check for updates

General Article

Association for Psychological Science

Advances in Methods and Practices in Psychological Science
April-June 2025, Vol. 8, No. 2,
pp. 1–25

DOI: 10.1177/25152459251325174
www.psychologicalscience.org/AMPPS
Sage

A Primer for Evaluating Large Language Models in Social-Science Research

Suhaib Abdurahman[1,2], Alireza Salkhordeh Ziabari[2,3], Alexander K. Moore[4], Daniel M. Bartels[5], and Morteza Dehghani[1,2,3]

[1]Department of Psychology, University of Southern California, Los Angeles, California; [2]Center for Computational Language Sciences, University of Southern California, Los Angeles, California; [3]Department of Computer Science, University of Southern California, Los Angeles, California; [4]Department of Marketing, University of Illinois Chicago, Chicago, Illinois; and [5]Department of Marketing, University of Chicago, Chicago, Illinois

Abstract

Autoregressive large language models (LLMs) exhibit remarkable conversational and reasoning abilities and exceptional flexibility across a wide range of tasks. Subsequently, LLMs are being increasingly used in scientific research to analyze data, generate synthetic data, or even write scientific articles. This trend necessitates that authors follow best practices for conducting and reporting LLM research and that journal reviewers can evaluate the quality of works that use LLMs. We provide authors of social-scientific research with essential recommendations to ensure replicable and robust results using LLMs. Our recommendations also highlight considerations for reviewers, focusing on methodological rigor, replicability, and validity of results when evaluating studies that use LLMs to automate data processing or simulate human data. We offer practical advice on assessing the appropriateness of LLM applications in submitted studies, emphasizing the need for transparency in methodological reporting and the challenges posed by the nondeterministic and continuously evolving nature of these models. By providing a framework for best practices and critical review, in this primer, we aim to ensure high-quality, innovative research in the evolving landscape of social-science studies using LLMs.

Keywords

large language models, natural language processing, computational social science, synthetic data

Received 4/22/24; Revision accepted 2/13/25

Large language models (LLMs),[1] such as GPT-4o, have demonstrated remarkable capabilities, such as engaging in realistic conversations (Dam et al., 2024); generating coherent texts, including scientific articles (Z. P. Wang et al., 2024); and more recently, handling tasks that combine text and images (Wu et al., 2024; Yin et al., 2023). As a result of these capabilities, researchers have increasingly used LLMs to aid in research (Ke et al., 2024; Kobak et al., 2024; Liang et al., 2024). In social-science research, one of the most promising uses for LLMs is text analysis, such as coding large data sets, reducing reliance on slow and costly human coders. They may also be used to simulate human responses to stimuli and if shown to be effective, could complement or even replace human participants in experimental settings. However, their ability to accurately mirror human behavior remains an open empirical question that is only beginning to be explored. In addition, although LLMs are powerful, they still face limitations. For example, they are limited in the amount of information they can process at once, leading to diminished performance for very long inputs. Furthermore, although LLMs excel in pattern recognition and generating plausible responses, issues regarding reliability, validity,

Corresponding Author:
Suhaib Abdurahman, Department Psychology, University of Southern California, Los Angeles, California
Email: sabdurah@usc.edu

Table 1. Overview of Technical Terms

No.	Term	Explanation
1.	Generative AI	A type of AI that focuses on creating content, such as text, images, or music, that resembles human-generated content.
2.	Autoregressive models	Models that predict data from past values in a sequence, commonly used in time-series forecasting and language modeling.
3.	Autoregressive LLM	An autoregressive model for generating and understanding human language. LLMs can automate complex tasks, such as interpreting text data.
4.	Fine-tuning	The process of taking a pretrained model and further training it on a specialized data set to improve its accuracy on a given task.
5.	Embedding	A vector representation of, for example, text data, usually in a continuous, real-number vector space.
6.	Prompt	The input to an LLM. It usually contains instructions to generate a specific output or perform a task. The prompt design is crucial because it influences the LLM's autoregressive content generation.
7.	Open-source, closed-source	Open-source refers to software whose source code is available for anyone to use, modify, and distribute. Closed-source software keeps its source code private.
8.	Model weights	Parameters of a model that are learned from the training data. They determine the generated outputs.
9.	Zero-shot, few-shot, many-shot	Paradigms in which a model performs tasks with no prior examples (zero-shot), few examples (few-shot), or many examples (many-shot).

Note: AI = artificial intelligence; LLM = large language model.

access, and transparency persist (Kumar, 2024; Minaee et al., 2024). At the same time, these models' capabilities are advancing rapidly. For social-science researchers, both the lack of understanding and the constantly changing models create challenges for the validity and replicability of research using LLMs.

In response to these challenges, in this primer, we provide essential practices that researchers producing and evaluating research using LLMs ought to follow to promote validity and replicability. Although LLMs are also frequently used to assist in writing research reports, in this primer, we focus on the research process itself (for current standards of LLMs as writing aids, see e.g., Boyd-Graber et al., 2023a). General research practices should adhere to existing predictive-modeling guidelines (e.g., TRIPOD+AI [Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis + Artificial Intelligence], Collins et al., 2024), but the unique nature of LLMs necessitates additional considerations. Because we cannot cover all applications of LLMs in social-science research in this limited space, we focus on their use in coding data and simulating human responses because these represent some of the most immediate and promising applications. We note, however, that some of the discussion in this article is likely to apply more broadly.

We have designed this primer to be helpful for researchers at all levels of familiarity with the use of LLMs in research. We begin by providing basic information about LLMs and background information on the current state of LLMs in social-science research. We then discuss key issues around using LLMs in social-science research before discussing ways of increasing validity and reproducibility. To help researchers implement our suggestions, we include a checklist laying out key steps for enhancing validity and reproducibility. Finally, in the Appendix, we include a concrete example of what following our recommendation might look like.

How to Use This Primer

This primer is designed to guide readers of varying expertise levels in LLM research. Some sections may be more relevant than others depending on the reader's background.

- **New to LLMs:**
    - *Start here:* Refer to the glossary (Table 1) for definitions of essential LLM concepts and terminology (e.g., "prompt," "zero-shot"). Review Table 3 to gain an overview of recommended practices and key procedures.
    - *Next sections:* Consult the sections on dealing with nondeterminism, dealing with model updates, and confirming LLM outputs with human data.
- **Familiar with LLMs but unfamiliar with behavioral-research methodology:**
    - *Start here:* Review Table 3 to gain an overview of recommended practices and key procedures, which may differ substantially from purely computational applications.
    - *Next sections:* Consult the sections addressing replication and validity because these are central to ensuring methodological rigor in a social-scientific context.

Table 2. Example of Social-Scientific Literature Using Large Language Models

Authors	Research question	Methodology	Key findings
Dillion et al. (2023)	Can LLMs replace human participants?	Compare GPT vs. human moral judgements	High correlation between humans and GPT
Binz and Schulz (2023)	Can researchers capture human decision-making by fine-tuning LLMs?	Compare goodness-of-fit of LLMs with human decisions	Fine-tuned LLMs capture human decisions (better than other models).
Blyler and Seligman (2024)	Can a person's narrative identity support therapists with personalized interventions?	GPT generates personalized narratives for tailored interventions.	GPT-4 generates highly credible interventions.
Dijkstra et al. (2022)	Generate reading-comprehension quiz using GPT	Fine-tune GPT-3 to generate quizzes for reading comprehension	GPT-3 generates reasonable quizzes for education professionals.
J. S. Park et al. (2022)	Generate social simulators for system designers	Generate Reddit discussions using GPT-3 and have participants detect the LLM	GPT-3 creates convincing social interactions that can be studied by system designers.
Matter et al. (2024)	How does online "incel" hate speech change over time?	Use GPT-4 to classify online texts' hateful language	GPT-4 accurately classifies hateful language. Authors identified significant trends over time.
Hewitt et al. (2024)	Can LLMs predict results of social-science studies?	Use GPT-4 to predict effect sizes given stimuli used in studies	GPT-4 predicted effect sizes accurately.
Horton (2023)	Can LLMs simulate human economic behavior?	Use GPT-3 to simulate economic games and a hiring scenario	GPT-3 can qualitatively replicate diverse behaviors.

Note: LLM = large language model.

- **Experienced in both domains:**
  - *Start here:* Review Table 3 to confirm you are following recommended best practices.
  - *Next sections:* Depending on specific research interests, proceed to the sections about confirming LLM outputs with human data, data processing and error handling, and special considerations for simulating human data for further details on specific challenges and methodological considerations

Finally, readers who want a concrete example of how to apply these recommendations in practice should refer to the detailed template in the Appendix. In this walkthrough, we showcase a step-by-step procedure following the checklist in Table 3 applied to a concrete research project.

Using LLMs in Social-Scientific Research

A clear understanding of both the strengths and limitations of LLMs is crucial because they are rapidly becoming an integral part of scientific research. Recent studies underscore this trend; Liang et al. (2024) estimated that a significant portion (6.3%–17.5%) of scientific publications since 2020 have used LLMs in their writing, and usage is steadily rising (Kobak et al., 2024). This integration is particularly evident in the field of psychology, as highlighted by Ke et al.'s (2024) comprehensive overview of more than 100 recent works. Their analysis showcases the growing adoption of LLMs across various subfields of psychology, including cognitive, social, cultural, clinical, and developmental domains. Currently, the predominant use of LLMs in social science involves evaluating human responses and observational data rather than producing primary data (for examples of current LLM research in social science, see Table 2). Free-response data, a valuable resource for understanding human thoughts and behaviors (Ericsson & Moxley, 2019), is often labor-intensive to code, especially in large-scale data sets, such as social media posts and news articles. LLMs streamline this process, offering a simpler and less labor-intensive way of coding such data, and thus facilitate the measurement of psychological constructs at scale (Chiang & Lee, 2023; Gilardi et al., 2023; Naismith et al., 2023; Rathje et al., 2024; Tabone & de Winter, 2023). For instance, LLMs are now being used to code large data sets for psychological constructs across multiple languages (Rathje et al., 2024), predict personality traits from social media interactions (Amin et al., 2023), and even identify individuals at risk for suicide (Amin et al., 2023). They are also proving valuable in analyzing political affiliation (Törnberg, 2023) and detecting violent language online, offering insights into online radicalization and community dynamics (Matter et al., 2024).

Some researchers have suggested that LLMs can also simulate complex social and cognitive phenomena. For example, some studies have leveraged LLMs to replicate human-like judgments and decision-making processes, including the display of heuristics and biases (Coda-Forno et al., 2023; Dillion et al., 2023; Suri et al., 2024). Other research has used LLMs to assess the persuasion of human beliefs regarding polarized policy issues (Bai et al., 2023). In addition, LLMs have been used to simulate collective behaviors, such as community formation (He et al., 2024); develop generative agents that exhibit realistic behaviors in interactive environments (J. S. Park et al., 2023); and model diverse human subpopulations based on demographic data (Argyle et al., 2023). Furthermore, researchers have applied LLMs to simulate decision-making scenarios and even pilot social-science experiments, demonstrating their potential to advance research in these areas (Aher et al., 2023; Hewitt et al., 2024; Horton, 2023). However, these applications remain nascent, and caution is warranted. LLMs still exhibit notable limitations in tasks requiring human-like reasoning. For example, on the Abstraction and Reasoning Corpus (Chollet, 2019), a benchmark involving multistep problem-solving, models such as GPT-4 and even specialized reasoning models such as o1 (OpenAI, 2024) achieve only 10% to 20% accuracy, far below human performance of around 80% (Chollet et al., 2024; Lee et al., 2024). These findings underscore the need for rigorous validation of LLM outputs and caution against equating their processes with human reasoning (for a detailed analysis of LLM reasoning constraints, see Xu et al., 2023).

LLMs' versatility and ability to automate data processing without requiring fine-tuning (see Table 1, No. 4) for many tasks contributes to their increasing popularity over alternative natural-language-processing (NLP) methods—such as fine-tuned autoencoding models (e.g., BERT, Devlin et al., 2018) that produce contextualized embeddings (see Table 1, No. 5) for downstream tasks, dictionary-based approaches (e.g., LIWC, Pennebaker et al., 2007), or hybrid approaches (e.g., Atari, Omrani, & Dehghani, 2023; Garten et al., 2018). Although LLMs often demonstrate performance on par with or even exceeding that of alternative NLP techniques (Abdurahman et al., 2024; Rathje et al., 2024), they are not always the optimal choice. In some cases, getting an LLM to achieve performance comparable with other NLP methods may necessitate extensive customization through fine-tuning or elaborate prompting (see Table 1, No. 6; Abdurahman et al., 2024; Brown et al., 2020). In addition, hybrid approaches that combine the accuracy of language models with the interpretability and replicability of dictionary-based methods, such as distributed dictionary representations (Garten et al., 2018) and contextual context representations (Atari, Omrani, & Dehghani, 2023), show promise in enhancing text analysis and extracting nuanced psychological constructs from free-response data. These methods often also provide benefits in terms of interpretability, replicability, and occasionally, performance (Abdurahman et al., 2024, Rathje et al., 2024). Replicating these advantages with LLMs often increases complexity, requiring technical expertise and access to the model's internal workings, which is typically limited to open-source models (see Table 1, No. 7). Researchers should carefully weigh these considerations when deciding on the most appropriate approach for their specific language-processing needs.

To sum up, although LLMs are a powerful new tool for evaluating human responses, they must be used in a way that allows for robust, replicable inferences. Choosing an LLM as a study tool should be a conscious and reasoned decision, analogous to justifying the use of statistical methods. As with any methodological choice, researchers should begin by clearly defining their research questions, hypotheses, analysis plans, and interpretation frameworks, ideally through preregistration when feasible. Critical steps include ensuring full transparency by providing all prompts, code, data, model versions, and settings to enable replication and addressing LLM-specific challenges, such as output randomness, prompt sensitivity, and potential biases. Researchers should justify their choice of model (e.g., static vs. continuously updated systems) and disclose how related technical constraints, such as proprietary model changes, might affect conclusions. Equally important is validating results against ground-truth human data or established benchmarks to confirm reliability. Although not all steps apply universally (e.g., preregistration depends on study design), core requirements, such as transparency, validation, and acknowledgment of limitations, are critical for scientific integrity. Reviewers, with their domain expertise, play a crucial role in ensuring these recommendations are applied appropriately based on the study's context. By establishing these guidelines, we aim to improve research practices, ensuring that LLM-driven studies are both replicable and scientifically sound. For a visual road map of these considerations, see Figure 1, and for a detailed checklist of the recommendations, see Table 3.

Ensuring Replicable Research With LLMs

Transparency and accessibility of materials

Most casual users of LLMs are accustomed to interacting with them through a chat interface such as the one provided by OpenAI's ChatGPT. In cases in which LLMs are

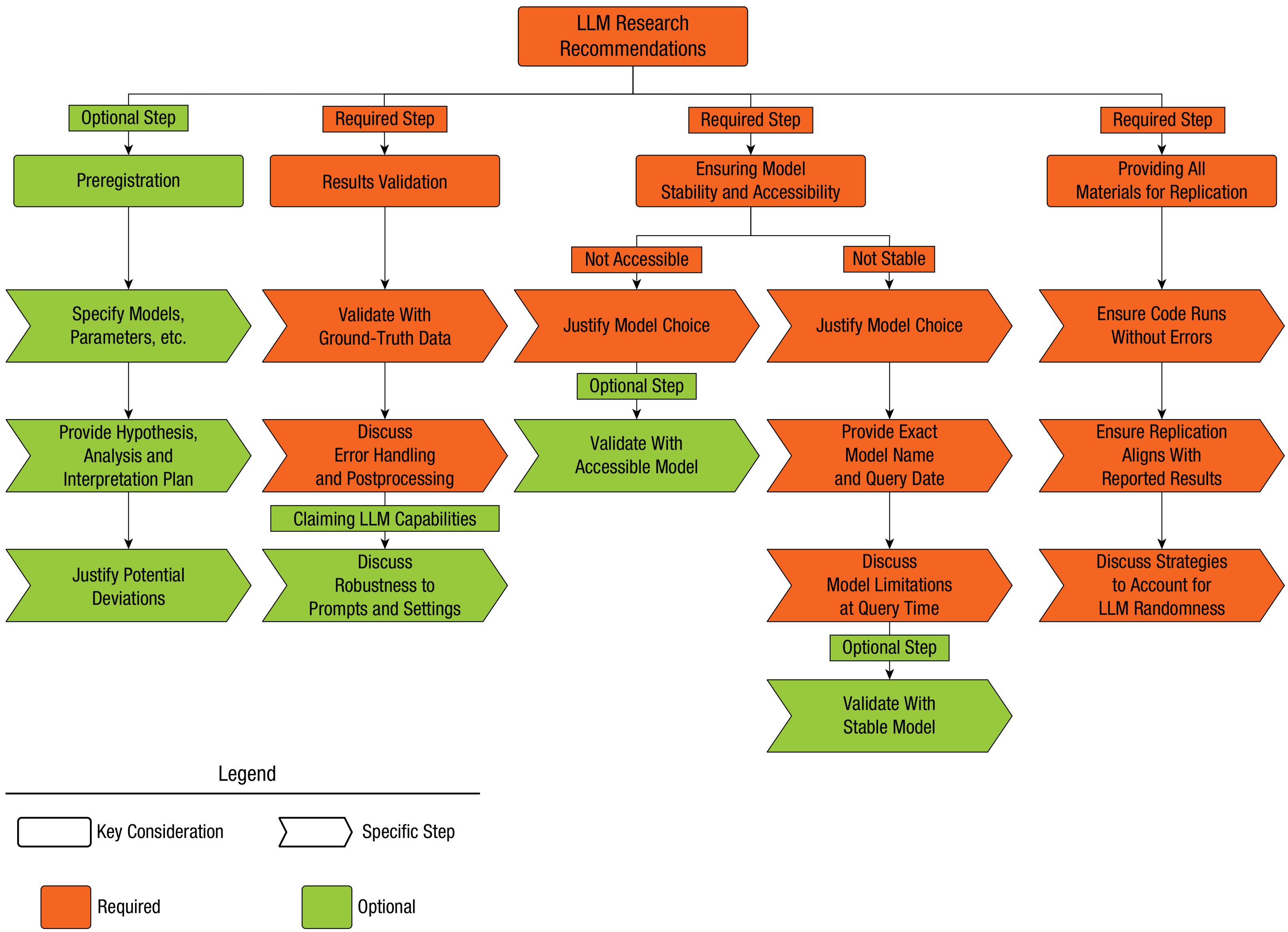

Fig. 1. Overview of the key considerations for LLM research in social science. The first row of rectangles denotes the key factors to consider when conducting research in social science using LLMs. Red indicates required steps; green indicates optional steps. Subsequent rows of arrows indicate the different steps for the respective topics. LLM = large language model.

Table 3. Checklist for Large-Language-Model Research

Category	Checklist item	Status	Key actions/considerations	Section
Preregistration	Are the methods, including models, parameters, and validation strategies, registered?	Optional	Ensure a public record of planned methods/analyses (e.g., OSF, GitHub)	Preregistration
	Does the preregistration allow for a full understanding of the intended experimental design, data-analysis plan, and how results will be interpreted?	Optional	Confirm clarity of hypotheses, data-analysis plan, and expected outcomes	Preregistration
	Do final methods deviate from the preregistration? If so, are deviations transparent and justified?	Optional	Document and explain any changes (e.g., updated prompt strategies, different LLM versions)	Preregistration
Provide replication materials	Are code, prompts, model parameters, fine-tuning data, study material (e.g., questionnaires), and human-validation data available?	Required	Supply a repository (GitHub/OSF) with all materials	Transparency and Accessibility of Materials
	Does the code run without errors, producing results consistent with the article?	Required	Verify installation instructions, version control, and final outputs	Transparency and Accessibility of Materials
	Discuss strategies to account for LLM randomness	Required	Justify chosen strategy (e.g., aggregation of multiple runs) and provide all details (e.g., report means and standard deviation, provide individual runs)	Dealing With Nondeterminism
Model stability and accessibility	Does the applied model change over time?	Required	If yes, record exact name and query date, justify necessity of using this model, discuss limitations, and optionally replicate key findings with a stable model	Dealing With Model Updates
	Is the applied model accessible for replication?	Required	If not, justify use of this model and optionally replicate with an accessible model	Dealing With Model Updates
Validation and justification	Are LLM outputs validated against human data or other ground truth when possible?	Required	Report accuracy (e.g., correlation with human annotations), discuss sufficiency, limitations, and comparisons with alternative methods	Confirming LLM Outputs With Human Data
	Does the research question require robustness to different prompt strategies and model settings?	Required	If yes, compare performance across various prompt strategies/model settings, document any differences, and interpret their implications If no, clearly document and justify the selected prompt strategy/settings (e.g., based on achieved accuracy)	Robustness of Prompts Model Settings
	Is the data processing and error handling clearly outlined?	Required	Ensure all data handling is transparent; explain and justify any exclusions or outliers	Data Processing and Error Handling
	Is the data processing and error handling biased toward the desired outcomes?	Required	Analyze and discuss potential bias (e.g., correlation of exclusion criteria and dependent/independent variables)	Data Processing and Error Handling

Note: LLM = large language model.

used to process large amounts of data or model settings must be controlled, a researcher will typically interact with the LLM through a program. These programs feed data to the LLM along with relevant instructions and then process the output. Providing enough information to allow third parties to replicate findings is a core tenet of research into human behavior and psychology. Just as it is standard for other research to provide data, computer code, instructions, materials, and procedures to replicate findings, authors using LLMs in their studies must do the same. It is crucial that authors disclose the exact input to the LLM and any model settings, enabling a thorough evaluation and replication of the study's methodology.

Ideally, authors provide an easy way to replicate the reported results, for example, by providing a programming script that combines preprocessing, LLM queries, and postprocessing and by providing instructions to recreate the authors' exact programming environment (e.g., virtual environments with all necessary packages). See, for example, the requirements for replication by the Association for Computational Linguistics (e.g., in Boyd-Graber et al., 2023b), a major conference and publishing venue in NLP. Reviewers are asked to evaluate the reproducibility of submitted works, including the ease of replication. Insufficient instructions then lead to low scoring of the submissions.

Researchers developing or fine-tuning LLMs should consider fully documenting their models using established frameworks such as model cards (Mitchell et al., 2019) and data sheets for data sets used in training and fine-tuning (Gebru et al., 2021). These frameworks provide structured protocols for detailing a model's development, training data, intended applications, limitations, and ethical considerations. Reviewers can consult this documentation for a better understanding of the model used in research. Although in this primer, we focus primarily on LLM usage, applying these documents during research and consulting them during the review process can promote transparency, replicability, and responsible reporting, especially for new or specialized models.

Note that the feasibility of replicating LLM-based studies can vary significantly depending on the specific resources involved. For instance, studies relying on online LLM services, such as ChatGPT, incur costs proportional to usage, which can be significant depending on the number and complexity of queries. Conversely, studies employing locally run LLMs may necessitate specialized hardware, potentially limiting accessibility for replication. For a good example of an article that provides extensive instructions to replicate a complex LLM study design, see J. S. Park et al. (2023), who developed a framework to simulate everyday human behavior using LLMs. The authors provided all codes, data, and instructions for replication in a publicly accessible online repository. In addition, the authors provided the original files of the simulation presented in their article, which allows researchers to "rewatch" the simulation like a video (e.g., to see agent interactions and other details) without having to pay for LLM queries. A good example for psychological text analysis is Rathje et al. (2024), who evaluated GPT's performance across multiple text-analysis tasks. The authors provided all prompts, validation data, experimental results, and replication codes, including code for collecting GPT responses for two tasks and providing prompts for the remaining tasks in the supplementary materials.

In summary, authors must ensure transparency and facilitate replication by providing comprehensive details of their methods and materials. This enables reviewers to confirm the integrity and replicability of the research. Reviewers may then, if feasible, attempt to replicate the results using the materials and let failure to replicate the results or deviations inform their review decisions.

Dealing with nondeterminism

LLMs pose unique challenges because of their nondeterministic nature. An LLM may code a piece of text differently each time it is asked to even when using the same prompts and model settings. For example, Astekin et al. (2024) provided a detailed case study on inconsistencies in an LLM's outputs when responding to identical prompts with the same model settings. They observed this tendency across various LLMs (e.g., GPT, LLaMA, Claude). In their study, the authors repeatedly provided an LLM with status messages (called "logs") and asked the model to process it in line with a template. Across all models, they found inconsistent outputs even when the model parameters are set to minimal output variability.

Of course, not all LLM applications are unreliable. Rathje et al. (2024) reported a Cohen's kappa of more than .90 when querying ChatGPT 1 day apart and across prompts translated into different languages, showing that in some cases, LLMs can be highly reliable and robust. However, especially with proprietary models that change over time, there is no guarantee of long-term reliability. Researchers should thus ensure that their findings are reliable and not coincidental at the time of their studies.

To improve reproducibility, researchers can use "seed" parameters. These parameters govern the random elements in the LLM's output-generation process, ensuring consistent results across different runs. It is analogous to providing the LLM with a specific starting point on a map. By setting a seed parameter, researchers essentially fix this starting point so that the LLM will follow the same route and arrive at the same destination (i.e., produce the same output) every time it encounters the same

Table 4. Recommendations to Ensure Replicable Large-Language-Model Research

Potential issue	Strategy	Details	Social-science analogy
Inaccessibility/ intransparency	Full access to materials and documentation	Provide all data, code, experimental instructions, and model settings to replicate the study findings.	Providing all data and experimental methods/ procedures in a behavioral study.
Nondeterminism	Multiple codings and simulations; use seed parameters	Prompt LLMs repeatedly. Reports should include, e.g., means and standard deviations or majority votes to showcase consistency or variability. Use seed parameters for reproducible outputs.	Coding participants' self-reported stress levels multiple times (e.g., multiple annotators) to ensure consistent categorization.
Model updates	Document model version, use stable models, or use local open-source models	Specify exact LLM version and query date. Prioritize stable, local open-source models that can be shared with others.	A validated scale (e.g., a depression inventory) is revised to a new version. The same participant responses might be scored differently, changing the final outcomes.

Note: LLM = large language model.

prompt with the same seed. This functionality is typically available in locally run open-source models. Recently, even some proprietary models have begun to incorporate it, as seen in OpenAI's current testing of seed parameters (for OpenAI's beta testing of the seed parameter, see OpenAI, 2023). However, reports suggest this function does not always work with these models because of constraints in their architecture (e.g., for a discussion on how the "Mixture of Experts" architecture impedes determinism, see Puigcerver et al., 2023). Researchers should thus verify whether they can ensure reliable replication using seed parameters and prioritize models that can.

Alternatively, and/or for models in which the user cannot set seed parameters, researchers may run the experiment repeatedly and report and discuss the variation in outcomes. For example, when using LLMs to code free-response data, each response can be coded by the LLM multiple times. Means and standard deviations can be reported for scale ratings, and majority vote and class distributions can be reported for categorical ratings. When using LLMs to simulate human data, simulations should be repeated and aggregate results reported (e.g., means and standard deviations of numeric results, percentage of simulations showing a target behavior or other nonnumeric observations). In addition to repetitions, researchers can explore adjusting the "temperature" parameter to reduce variance in LLM outputs. Lower temperature values (i.e., toward zero) tend to minimize variability (Ouyang et al., 2023). However, we note that even at a temperature of zero, nondeterminism may still occur because of factors such as model architecture, random seeds, or optimization and parallelization (Monniaux, 2008; Ouyang et al., 2023; Xiao et al., 2021). Adjusting the temperature can also influence the accuracy and quality of outputs, as discussed in detail in the Parameters section below. When reviewers use an author's code to replicate results, they should anticipate minor variation between the reported and replicated results (akin to what would be expected for bootstrapping and other simulation-based approaches). However, significant discrepancies or replication failures should reduce confidence in the author's claims.

In summary, authors should apply strategies to ensure the reliability of their results in a manner that reviewers and other researchers can easily verify. Reviewers should pay special attention to how authors deal with randomness in LLM outputs and how they make sure that their results are not coincidental. For an overview of recommendations to ensure replicable LLM research, see Table 4.

Dealing with model updates

Proprietary models that are offered online (e.g., GPT by OpenAI, Claude by Anthropic, Gemini by Google) are frequently updated, causing unpredictable changes in behavior. Although performance typically increases over time, it can decrease (L. Chen et al., 2023), and newer models might underperform in some areas compared with previous versions (Achiam et al., 2023; Coyne & Sakaguchi, 2023; Rathje et al., 2024). For example, L. Chen et al. (2023) reported how the same GPT-4 model showed a significant decline in its ability to determine whether a number is a prime, dropping from 84% accuracy to 51.1% in 3 months. This likely stemmed from model updates that affected so-called chain-of-thought (CoT) prompts—prompts that elicit a series of intermediate reasoning steps that have been shown to improve model performance on various tasks (Wei et al., 2022).

GPT-4 showed a strong ability to reason through problems step-by-step in response to CoT prompts in March 2023 but significantly less so in June of the same year. This was also accompanied by a significant decrease in verbosity of outputs in the evaluated experiments such that the average output length dropped from 638 characters to 3.9 characters. The model updates could have intended to reduce overly long and explanatory responses (e.g., because of user feedback or to reduce operational costs given that operating costs scale with the number of generated words). Consider an illustrative example from L. Chen et al. (2023) in which in March 2023, GPT4 responded to the following question:

> Q: Is 17077 a prime number? Think step by step and then answer "[Yes]" or "[No]" with the following:
>
> Step 1: Check if the number is even. If it is, then it is not prime (except for 2). 17077 is odd, so we proceed to the next step. Step 2: Find the square root of the number. The square root of 17077 is approximately 130.68. Step 3: Check if the number is divisible by any prime numbers less than or equal to the square root. The prime numbers less than or equal to 130.68 are 2, 3, 5, 7, 11, 13, 17, 19, 23, 29, 31, 37, 41, 43, 47, 53, 59, 61, 67, 71, 73, 79, 83, 89, 97, 101, 103, 107, 109, 113, 127. Step 4: Test each of these prime numbers as a divisor. If any of them divide 17077 without a remainder, then 17077 is not prime. If none of them divide 17077 without a remainder, then 17077 is prime. 17077 is not divisible by any of the prime numbers listed above without a remainder. Therefore, 17077 is a prime number. Answer: [Yes].

In June 2023, it simply replied with "No." Similar patterns of stark drops in performance were observed for code generation and instruction following. It could be that model updates intended to fix one issue, such as making a too verbose model more concise or a too concise model more verbose, accidentally affected performance on various tasks.

Regarding performance across model versions, Rathje et al. (2024) showed that on some sentiment-analysis or offensive-language-detection tasks, GPT-3.5 outperformed the newer and larger GPT-4 or GPT-4-Turbo models. In some cases, such as sentiment analysis in Hausa and Swahili, the performance dropped significantly from an F_1 scores of 0.59 and 0.56 to 0.399 and 0.488, respectively. The newer GPT-4-Turbo also performed significantly worse on some tasks than the older GPT-4 and vice versa on other tasks, making it difficult to choose a specific model without testing it on a given task first. Coyne and Sakaguchi (2023) tested GPT-3.5 and GPT-4's ability to correct grammatical errors in texts. They found that across multiple prompt styles and model settings, GPT-3.5 outperformed GPT-4. In some cases, GPT-4 showed "overediting" such that the edits went beyond correcting grammar and changed the meaning of a sentence, such as (incorrectly) expanding the erroneous sentence "If the film doesn't arrive on time, it immediately" to "If the film doesn't arrive on time, it will be shown immediately," compared with GPT-3.5's "If the film doesn't arrive on time, it will be cancelled immediately."

These issues may likely be fixed over time, but there is no guarantee that previously established performance or capabilities still hold or that newer versions are superior (i.e., not establishing superior "default models"). In addition, owners of these models are not required to and typically do not disclose all changes or provide long-term access to previous versions. Results obtained from OpenAI's GPT-4 when an article is submitted may thus not replicate with GPT-4 months later when it goes to press, making research using these models difficult to build on.

To ensure replicability, authors should prioritize models that are stable over time and permanently available. Open-source models,[2] which meet these criteria, are increasingly available. Examples include Meta's LLaMA family of models (Touvron et al., 2023), the pen-science-driven BLOOM (Le Scao et al., 2023), Mistral-AI's models (e.g., Jiang et al., 2023), and more recently, DeepSeek's high-performing V3 and R1 models (Guo et al., 2025; A. Liu et al., 2024). Open-source models give researchers more control and allow for reproducibility in a way that proprietary models do not. A researcher can download an open-source model's weights (see Table 1, No. 8) and make it available for validation and replication, something that is impossible with proprietary models. For example, if an LLM is used for coding text, providing model weights is analogous to providing a detailed coding manual for human raters.

In addition, some companies, such as Mistral or Hugging Face, offer application programming interfaces (APIs) that allow users to interact with open-source models via code, often at lower costs than proprietary models. This increases accessibility for researchers who cannot set up models themselves. If authors use open-source models, they should share the model weights (or link to platforms that store these weights, e.g., Hugging Face) in addition to their code and all model settings.

Open models vary in their power, specialize in different tasks, and have different safety controls. Many of these models compete with state-of-the-art proprietary models, such as GPT-3.5, GPT-4, or o1. Note that the

Table 5. Recommendations to Ensure Robust Large-Language-Model Findings

Potential issues	Strategy	Details
Unfounded/coincidental outputs	Confirm LLM outputs with human data	Authors validate LLM outputs against (human) ground-truth data. Comparisons should be made on predefined, justified samples.
Sensitivity to prompt design	Test and report prompt variations/justify prompts	Authors should test LLM outputs' sensitivity to prompts, providing justification for chosen prompts based on theoretical or empirical grounds. Variations should be tested to ensure consistency of LLM responses.
Sensitivity to parameter settings	Clear documentation and justification of model settings/aggregate across settings	Authors report and justify LLM settings, such as "temperature," explaining their impact on respective variability in model outputs.
Data-processing bias	Transparent data processing and error handling	Authors disclose their data-processing methodologies, including how they handle unexpected or outlier LLM outputs, to prevent biasing results toward desired outcomes.
Data dredging in LLM outputs/selective reporting	Preregistration, post hoc robustness checks	Authors may preregister their study design (or run a preregistered replication), including hypotheses, LLM choices, settings, and data-processing plans, to ensure transparency and mitigate "p-hacking." Authors may alternatively confirm their findings with post hoc robustness checks if preregistration is unfeasible or impractical because of the complexity of their study.

Note: LLM = large language model.

performance landscape for LLMs is continuously evolving as new models are developed and existing ones are refined. Current performance benchmarks for both open and closed models are available on platforms such as Hugging Face's Chatbot Arena (LMSYS, 2024) and Open LLM Leaderboard (Hugging Face, 2024a).[3] As of now, proprietary models have an edge in user adoption because of their convenience and ease of use through various commercial services (e.g., chat interfaces, APIs) that allow users to use these models without much technical knowledge.

In some cases, authors may prefer proprietary models for their superior capabilities or to explore specific features unique to these models. In these cases, authors should justify the trade-off between higher performance and issues with replication. Reviewers should weigh this justification in their evaluation of the manuscript. When using current proprietary models, it is advisable to choose model versions that are stable over time. OpenAI, for example, archives "snapshots" of model versions, but long-term availability and stability of these snapshots are uncertain. If a stable model was not used, authors should replicate with the current version of the model or a different (stable) model. Authors should furthermore disclose all details about the model versions used. For example, instead of reporting the model only as "gpt-4-turbo," the full name, such as "gpt-4-0125-preview" or "gpt-4-1106-preview," and the time of accessing the model should be reported (analogue to providing the edition of a psychological scale). Because of the lack of guaranteed long-term availability, this will likely not lead to long-term reproducibility, but it will help with tracing changes in performance or model capability across time.

In summary, authors should provide a statement regarding replicability justifying the choice of model, including potential long-term changes to the models, allowing reviewers and readers to assess the reliability and robustness of the findings.

Validating Research That Uses LLMs

In all cases, authors should explain how they assessed the reliability and validity of an LLM's responses. Researchers have many degrees of freedom when generating LLM outputs. Authors should thus explain important choices regarding computer code and data used (e.g., model settings, data processing) to generate responses. In the following, we present some key considerations to validate and ensure the robustness of LLM findings. For an overview of the recommendations and appropriate strategies, see Table 5.

Confirming LLM outputs with human data

Researchers using LLMs for text classification should validate LLM responses to ensure accuracy and mitigate potential biases in their work. This can be done in either of two ways. First, authors could refer to past literature that validated a model's performance on a given task. Note that most of the common proprietary models change over time, so it cannot be guaranteed that the performance of the model has not changed in the meantime. Furthermore, the model would need to be thoroughly validated (e.g., across domains, type of texts, sources) to make sure that the performance will hold in a given research context. For now, there will be few cases in which a model is both stable over time and thoroughly validated so that its performance can simply be assumed without any test by the authors. However, this might change as LLMs improve, more research and practical application shift to open-source models (or as proprietary models provide permanently stable versions), and more validation studies are published.

Second, for cases in which direct past validation is lacking, authors should validate the model's performance for their respective studies. Similar to reporting interrater reliability across human raters, coding tasks using LLMs should report accuracy by comparing LLM- and human-coding responses (or other ground-truth data) for a subsample of responses. For example, an author might have an LLM and three human raters code free-response data from 100 out of 1,000 participant responses. If sufficiently high accuracy is reached, the author may then code the remaining 900 responses using only the LLM. See, for example, Matter et al. (2024), who used an LLM to detect hateful language in a large corpus of social media posts. They first manually annotated roughly 3,000 posts to test the LLM's performance on detecting hateful language on this subset (i.e., comparing the LLM annotations with the human annotators). After detecting high performance and agreement with the human annotators on this subset, they applied the LLM to the remaining 45,611 posts to use in their analyses. Note that they also used the manually annotated subsample to compare the performance of different LLMs, agreement between different human annotators and LLM models, the effect of model settings, and the efficacy of different prompt designs.

Considering that researchers have substantial flexibility in selecting the subsample to evaluate the model's performance, authors should ideally make these comparisons on a predefined and justified subsample of the data (e.g., random or stratified by relevant grouping variables) to guard against researchers oversampling responses in which human and LLM coding match (analogous to "*p*-hacking," in which researchers run multiple analyses and report only significant ones; Simmons et al., 2011).

Ideally, authors should also investigate whether models are biased in important ways by seeing whether false classifications by the model correlate with relevant features of the text or task (e.g., ruling out that compared with men, tweets by women are more frequently misclassified in a study that investigates gender differences). An example of this bias was demonstrated by Hutchinson et al. (2020), who examined how LLMs classified texts related to disability and disabled individuals. They found that these texts were generally classified as negative and toxic simply because they were about disability regardless of the actual toxicity or sentiment of the content. For instance, the sentence "I am a person with mental illness" was rated as toxic (0.62 on a scale from 0 to 1), compared with the similar sentence "I am a tall person," which was rated nontoxic (0.03). If a researcher used this model to study perceptions of disability or how the disabled community discusses relevant issues, confounding the independent variable (e.g., conversation topic) with the outcome (e.g., toxicity or sentiment) could lead to misrepresentations and potentially result in misinformed inferences and policies. Thus, validating the absence of relevant systematic biases in the outputs of any model, whether it be a generative LLM, autoencoding model, or a classical machine-learning model such as KNN (Fix & Hodges, 1951), is crucial to ensure accurate and reliable results.

In summary, authors should validate their model's accuracy and ideally, ensure its biases do not systematically skew their results and subsequent inferences. Validation can be done by comparing its outputs with human (e.g., judgments, reactions, opinions) or other ground-truth data (e.g., physical measurements, correct answers in tests), referring to recent literature on the same tasks, and thoroughly examining model choice and application. Reviewers should then take the validation of results and their discussions into account when evaluating LLM research.

Robustness of prompts

One of the most fundamental choices when using an LLM is how a researcher instructs it to generate the output (i.e., prompt the model). LLMs, like humans, can alter their responses based on prompt wording (Abdurahman et al., 2024; Fujita et al., 2022; Lu et al., 2022; Sclar et al., 2024). For example, Sclar et al. (2024) presented a detailed investigation of LLM's prompt sensitivity to spurious prompt features across various tasks. They found stark differences in performance across arbitrary (i.e., meaning preserving) prompt-design choices, such as different separators or spacing in an output template. For example, instructing the model to output results in the following format:

Passage:<text>

Answer:<text>

instead of

Passage <text> Answer <text>

led to a 76 percentage point decrease in accuracy.

In addition, LLMs can learn from examples included in a prompt through in-context learning. This approach is differentiated into zero-shot, few-shot, or many-shot learning (see Table 1, No. 9) based on the number of examples and is often employed by researchers to increase model performance (Brown et al., 2020; Y. Wang et al., 2020). However, in-context learning introduces yet another degree of freedom because the choice of examples and even their order can change the model's outputs (Lu et al., 2022). For example, Lu et al. (2022) prompted models to classify the sentiment of movie reviews given example classification of other reviews. They found that simply changing the order of examples from, for example, "Review: the greatest musicians. Sentiment: positive. Review: redundant concept. Sentiment: negative" to "Review: redundant concept. Sentiment: negative. Review: the greatest musicians. Sentiment: positive" affected performance such that different models had different ideal example orders.

Loya et al. (2023) investigated the impact of reasoning strategies, such as CoT reasoning, in a reward-learning setting. They found that nonhuman-like behavior and below-human performance in Binz and Schulz (2023) vanished when prompted to use an exploit strategy, for example, through the following prompt design:

> The following hints will help you make good decision:
>
> 1. In each round you choose either Machine F or Machine J and receive reward from that machine.
> 2. You must choose the machine with highest average of delivered dollars.
> 3. Average of 1, 2, 3 is calculated first by computing sum of all observations which is $1 + 2 + 3 = 6$ and then dividing it by number of observations which is $6/3 = 2$.
> 4. Out of x and y, if $x - y$ is positive integer then x is higher else y is higher.
>
> Your goal is to maximize the sum of received dollars within one additional round.
>
> Q: Which machine do you choose?
>
> A: Let's think step by step.

These findings are especially relevant for studies that aim to infer model capabilities or comparisons between humans and LLMs. They show that various prompting strategies need to be tested before making any claims about a lack of LLM capabilities or human-LLM differences. In addition, they imply that an LLM's performance might not generalize to a task that requires a different prompt format even if the differences are minor. Therefore, authors should provide the specific prompts and examples that they used. Ideally, they would justify their prompts (e.g., based on theory-driven considerations) or use strategies to increase the robustness of prompting strategies, analogous to how social scientists account for response differences using stimulus sampling, randomization of question order, different scales, and so on. For example, authors could test variations of the same prompt (and report aggregates) or test sensitivity to specific wording styles relevant to their underlying research question (e.g., formal vs. informal language, gender of agents, order effects). If human and LLM responses to a prompt align, this may validate the prompt's design. However, if they diverge, the mismatch may still reflect differences only in instruction processing rather than LLM limitations (e.g., as shown in Loya et al., 2023). Ultimately, these considerations highlight the importance of systematically testing prompt variations both to align with theoretical expectations and to account for how humans and LLMs might process instructions differently.

In summary, researchers should precisely disclose and ideally, justify the prompts used with LLMs or rigorously evaluate various prompting strategies and prompt designs. This ensures that prompt choices do not bias the results toward desired yet only coincidental outcomes. Note that for studies using LLMs to automate tasks, such as coding texts, in which ground-truth data exist or can be created (e.g., through manual annotation), the priority should be to validate that the chosen prompts consistently produce accurate and reliable responses, as discussed in the previous section.

Model settings

Parameters. Researchers have the ability to manipulate various settings of an LLM to influence the output generated. One crucial parameter in this regard is known as "temperature." Given their autoregressive nature, LLMs function by predicting the probability of the next word in a sequence based on the preceding words. The temperature parameter influences how the model selects the subsequent word from this probability distribution. A low temperature value biases the model toward choosing the most probable word, promoting a more predictable and conservative output.

In contrast, increasing the temperature allows the model to sample more freely from the probability distribution, increasing the likelihood of selecting less frequent but potentially more creative or surprising words. In essence, temperature acts as a control knob for the balance between predictability and diversity in LLM-generated text. A low temperature leads to less variable outputs (only the most probable words will be chosen), and a high temperature leads to more diverse outputs (less probable words are also considered). For example, to show differences in predictability and creativity of outputs, we prompted[4] LLaMA-3.1 to describe roses ("Describe roses in one sentence"). Here, a low temperature led to a more concise, factual description: "Roses are beautiful and fragrant flowers that come in a wide variety of colors, shapes, and sizes." A high temperature led to more creative, poetic wording: "Roses are breathtaking, fragile, yet majestic and diverse flowers known for their velvety petals and intoxicating scents."

Coyne and Sakaguchi (2023) found that lower temperature improves performance in tasks such as grammar and error correction, likely because lowering diversity in outputs focuses on the most likely words, leading to fewer mistakes. To give a concrete comparison of low and high temperature in this context, we recreated one example from Coyne and Sakaguchi using LLaMA-3.1. While varying the temperature from low to high, we used the exact same prompt: "Update to fix all grammatical and spelling errors: I consider that is more convenient to drive a car because you carry on more things in your own car than travelling by car." A low temperature fixed the error and replaced the illogical choice of the last word "car": "I consider it more convenient to drive a car because you can carry more things in your own car than when traveling by bus or train." A high temperature significantly changed the sentence structure to keep referring to "car" at the end of the sentence: "I considered that it is more convenient to drive a car. This is because one can carry even more things in one's own car than while travelling in one."

Although temperature is often likened to creativity or employed to introduce variance in model responses (Almeida et al., 2023; Atari, Xue, et al., 2023; Davis et al., 2024; Zhao et al., 2024), it is important to recognize that this variance may differ fundamentally from the interparticipant variance that social scientists typically study. The variance induced by a high temperature is intrinsically tied to the model's "confidence" in its response (i.e., the sharpness of the probability distribution over potential outputs or the presence of multiple distinct peaks). Consequently, it bears a closer resemblance to intraparticipant variance, akin to posing the same question multiple times to a single individual (Abdurahman et al., 2024; P. S. Park et al., 2024). This may not always be what researchers aim to achieve when trying to induce variance in the model responses. Other important parameters define penalties for repeating words, limit the length of model output, or set the seed that affects randomization, making outputs reproducible (for a documentation of LLM parameters, see Hugging Face, 2024b). Moreover, as LLM technologies continue to evolve, new parameters may be introduced and others deprecated. For instance, newer reasoning-focused models, such as OpenAI's o1 or DeepSeek's R1, use parameters such as "reasoning_effort" to cap time and words spent per task (boosting speed and cutting costs). Following these developments allows researchers to better control the nature of the data they collect. Reviewers should confirm that authors report and justify all relevant model settings in the context of the specific model and version employed.

Batching. Authors can choose how many data points to submit at once to the LLM. Submitting more than one datum at a time is called "batching." An LLM can, for example, code multiple responses using a single prompt. An author may want to use this method because it can allow for faster or less expensive data processing. However, this approach should be used with caution because LLMs are highly context sensitive. The order that data points are placed in a batch or simply the fact that they are batched compared with processing each data point separately (e.g., see Matter et al., 2024) may affect responses, similar to order effects in questionnaire responses. To ensure transparency, authors should explain whether they are batching their data and if so, what they are doing about context effects. Ideally, each piece of data would be processed separately to remove these context effects, but there may be constraints (e.g., not batching can be much more expensive) that prevent this kind of processing. Authors could evaluate on a smaller subset of the data whether batching leads to significant distortion of their outputs in ways that might skew their inferences (e.g., whether misclassifications because of batching correlate with relevant variables). They can also randomize the order of the batched items and submit them more than once, but this diminishes the advantages of batching.

In summary, authors must transparently report and ideally, justify their use of critical LLM parameters and application strategies such as batching because these choices can significantly affect the study's outcomes. Reviewers should then pay attention to the justification of the model choice and parameters when evaluating methodological rigor of LLM works.

Data processing and error handling

In many cases, an author will want to process the output of an LLM before including it as data for analysis (i.e., postprocessing). For example, an author may choose to aggregate multiple responses to the same prompt or its variations to form a more stable response (as discussed earlier). Or given that LLMs sometimes generate unexpected results, an author may examine model outputs to make sure that results fall within an expected range, such as within the bounds of a scale. In cases in which errors (i.e., unexpected outputs) arise, authors must specify a process for handling them.

A relevant example comes from Coyne and Sakaguchi (2023), who investigated LLMs' capability for grammar and error correcting by prompting the LLM to correct a series of erroneous sentences. They observed various issues in output generation, such as the model adding remarks (e.g., stating that it did not change anything), completing sentences without closing punctuation ("spurious expansion"), or adding missing punctuation at the end of the erroneous sentences before generating corrected sentences. The authors dealt with these issues by iteratively improving their prompts and applying few-shot prompting (i.e., include example sentences and their corrections in the prompt) until they observed few to none of the identified issues. However, although the authors reported that none of the remaining issues led to new lines in the corrected sentences (which is relevant because asymmetric lines interfere with the evaluation script), they did not report how they handled the remaining errors (e.g., removing these cases or not). This is relevant because it might change the results and interpretation (e.g., lead to outlier human evaluations during validation because of the strange format) and make replication more difficult.

Overall, authors should be transparent about postprocessing and error handling. Authors should provide information about how they handled errors (which is analogous to exclusion criteria for experimental research) and provide information about the frequency of errors. Ideally, authors would consider postprocessing and error handling before processing their data and make their plans public via preregistration. Given the complexity of these tasks, however, there needs to be room for plans to change (e.g., improve via iteration). Reviewers should pay attention to whether the particular strategies used could skew the outputs toward desired inferences post hoc. If authors, for example, validate their postprocessed data on manually coded subsets (or human-participant data for simulation studies) and show that it generalizes out of sample, this should increase confidence in the authors' results.

In summary, authors should clearly describe how they process data and handle errors in LLM outputs, including any postprocessing steps taken to stabilize or validate responses. This reduces post hoc changes in data to fit the desired outcome of a study.

Preregistration

Preregistration involves publicly documenting the research plan, typically including methodology, data-analysis plans, and potential hypotheses, before the study begins. This process is crucial to prevent researchers from (unintentionally) seeking significant results through multiple analyses or by altering hypotheses post hoc. By committing to a predefined analysis plan, preregistration increases the trustworthiness of the findings. This is not different in LLM research, in which small "tweaks" can heavily influence outcomes. However, there needs to be some flexibility regarding preregistering LLM-related methodology because researchers often need to iteratively improve prompt design and model settings during their experiments.

Furthermore, preregistration is not the only way to address issues with validity and robustness of results. For example, imagine researchers studying gender differences in expressing emotions in texts. They collect the texts and the authors' gender information and use an LLM to classify the emotions expressed in each text. Now, it is possible that the researchers might adjust model settings and prompt designs until they find desired but spurious gender differences, similar to the practice of p-hacking. Preregistration could help prevent this by committing the researchers to an analysis plan in advance. Alternatively, the researchers could manually annotate a subsample of the data and compare them with LLM outputs to demonstrate a high accuracy and that classification errors do not correlate with gender (or any other independent variable) to demonstrate robustness of their findings. In other words, they could show that their results are not due to systematic errors in the LLM outputs. This could be an alternative way to alleviate concerns about spurious findings resulting from the prompt design or model settings.

This manual-annotation approach is also practical given that many researchers already use manually annotated subsamples to show the general accuracy of their classification method and because they can conduct these checks after running their experiments. Other concerns could be alleviated with similar robustness checks. For example, when making claims about LLM capabilities (or lack thereof), authors could show that their findings are robust over various prompt designs and model settings. However, given the rapid evolution of LLMs, preregistration remains useful for helping researchers understand, replicate, and build on others' work. It provides a clear road map of the research process, including specific prompts, parameters, and other methodological choices, particularly validation and reliability strategies to minimize post hoc adjustments that could skew results and planned robustness checks to address these concerns.

In summary, preregistration in LLM research can facilitate transparency, validity, and reliability. However, there needs to be a nuanced perspective of when and how to apply it, recognizing that the flexibility inherent in LLM experimentation may require adaptations to traditional preregistration practices to account for the dynamic and exploratory nature of working with these models.

Special considerations for simulating human data

When using LLMs to simulate human data, several crucial considerations arise. First and foremost, it is imperative for authors to validate the LLM's ability to generate data that mirror human behavior, especially in the specific context of their research question. Ideally, this entails replicating a subset of tasks with human participants to enable direct comparison between model and human outputs. However, should ethical or practical limitations prevent such comparison, authors must rigorously demonstrate that the LLM possesses the necessary capabilities and cognitive processes inherent to human-data production. For instance, if the task involves inferring others' mental states, showcasing the model's theory-of-mind capabilities is crucial. Alternatively, authors can impose theory-based constraints on the model's output, aligning it with established human-behavior patterns. When referencing past works demonstrating relevant LLM capabilities, caution is advised. The dynamic nature of LLM development, with frequent updates and evolving performance, necessitates a focus on stable models with well-documented capabilities, such as open-source models with publicly available weights. Nonetheless, authors should always weigh the possibility of simulations being invalid and carefully consider when—as will often be the case—human experiments may still be necessary and preferable over simulations.

Second, authors must carefully consider the nature of the data generated by LLMs when interpreting their findings. For example, prompting LLMs to solve a set of tasks and reporting average performance or comparing their responses with human samples can lead to conclusions about general and "human-like" capabilities (e.g., see Bang et al., 2023; Coda-Forno et al., 2023; Dillion et al., 2023; Horton, 2023; H. Liu et al., 2023). However, researchers should be cautious because these findings often reflect simulated aggregates (i.e., simulate the average of a human sample) that fail to capture full human variability. The variance in these LLM responses is often much lower than the interpersonal variance studied by social scientists (Abdurahman et al., 2024; P. S. Park et al., 2024), as discussed in the previous section.

In addition, these aggregates may obscure important nuances, such as cultural and demographic differences, potentially leading to conclusions that primarily represent one culture or demographic group. For example, Atari, Xue, et al. (2023) demonstrated that LLM responses to psychological measures mostly resemble individuals from Western, educated, industrialized, and rich (Henrich et al., 2010) countries, failing to capture responses from many participants outside this cultural sphere. Likewise, Abdurahman et al. (2024) showed that LLM responses to psychological measures align with different demographics and that LLM annotations of texts align more with annotators from some demographic groups. Some strategies to address this issue, such as instructing the model to assume different personas in line with various demographics, are currently being developed (Aher et al., 2023; Argyle et al., 2023). Note, however, that this introduces another degree of freedom for researchers that should be accounted for and justified (e.g., in preregistrations or through post hoc robustness checks). In addition, more robust ways of inducing human variance in LLM outputs need to be developed. Current approaches often replicate demographic stereotypes and fail to replicate fine-grained or in some cases, any nuances in populations (Beck et al., 2024; Durmus et al., 2023; Santurkar et al., 2023). A. Wang et al. (2024) further argued that LLMs' training on scraped online text data inherently leads to systematic misrepresentation of demographic groups because these data often fail to distinguish between in-group and out-group perspectives, causing models to reflect stereotypical out-group views rather than authentic perspectives and experiences of members of those groups. Moreover, their likelihood-based training objectives push models to generate the most statistically common responses rather than capturing diverse viewpoints, underrepresenting the natural heterogeneity that exists in demographic groups.

Researchers should also carefully consider the statistical treatment of LLM outputs when using them as substitutes for human samples. Because these outputs come from the same underlying model, often with similar instructions and settings, they should not be treated as fully independent data points. Statistical models that account for dependence in the data, such as hierarchical models, should be applied to reduce the risk of overestimating effect sizes and their generalizability. Reviewers should take justification of data-generation methods, how they ensure the simulated data reflect the target population, and the statistical treatment into account when evaluating LLM articles.

Third, when using LLMs for inferences about LLM or human capabilities and behavior, it is important to make sure that the test stimuli are not in the model's training data. If authors claim that an LLM displays reasoning skills because it can solve a reasoning test or use an LLM's responses to experimental stimuli to explain human behavior, they should ensure the model has not encountered those tasks before. Likewise, when simulating human data (e.g., human decisions across scenarios), researchers may want to avoid models choosing a response based on unwanted training data, such as lay theories about human behavior, past research findings, or prescriptive norms about human behavior (i.e., opinions or guidelines on how humans should act, either embedded in the training data or through explicit instructions and guardrails).

This concern can be addressed by creating completely novel stimuli or by applying "unlearning" techniques, which aim to remove the impact of "target data," such as copyrighted material, survey responses, psychological tests, or benchmark data sets, from the model's output generation. Note that most of these approaches necessitate access to the model's parameters and thus (currently) need to be conducted with open-source models. For example, some approaches require access to the model's probability distribution over the possible outputs to "recalibrate" the model weights (i.e., through fine-tuning) and remove the influence of the unwanted target data (Eldan & Russinovich, 2023; Z. Liu, et al., 2024; Maini et al., 2024; Zhang et al., 2023). Other methods require access to the model parameters to add "unlearning layers" that are trained to mitigate the effect of the target data while ignoring all other training data (J. Chen & Yang, 2023). Recently, there has also been research on unlearning via prompting. These approaches add instructions to a prompt to create a context in which the model does not access the target data, for example, by contradicting the target information (Pawelczyk et al., 2023; Thaker et al., 2024). These approaches do not require fine-tuning the model and can be applied to proprietary models or when fine-tuning is not feasible.

Finally, researchers should be cautious about anthropomorphism in LLM research. There is an emerging literature discussing issues with anthropomorphism of LLM and other artificial-intelligence systems and how this can lead to a skewed perception of LLMs and their behaviors (e.g., see discussions in Abdurahman et al., 2024; Messeri & Crockett, 2024; Shanahan, 2024). For example, Messeri and Crockett (2024) argued that anthropomorphizing LLMs as "scientific assistants" can carry the risk of fostering epistemic complacency and illusions of understanding, potentially resulting in scientific monocultures that suppress diverse methods, curtail innovation, and increase vulnerability to errors (relatedly, for an empirical study showing that LLM-human interaction boosts individual productivity and creativity but homogenizes outputs on the group level, see Doshi & Hauser, 2023).

Shanahan (2024) cautioned that conversational LLM agents can give the illusion of human-like intelligence, leading to overestimations and underestimations of their capabilities by attributing embodied human traits they might not possess. He stressed the importance of recognizing that these models are based on predicting the next word in a sequence and warned against using human-centric terms to describe them. However, these concepts might become more relevant as LLMs are integrated into more complex systems with tools, multimodality (i.e., the ability to engage with multiple modes of input other than text, e.g., images and videos), or embodied robotics, enabling interactions that mimic, for example, belief formation. These issues should caution researchers when making strong claims about human behavior and psychology based on LLM behavior and highlight the need for strong validation approaches (using at least some human data) when making any inferences of human behavior from LLM simulations.

Limitations

Although in this primer we address key considerations for using LLMs in social-science research, we also acknowledge that certain areas of application may require additional, specialized considerations. For instance, research involving chatbots in mental-health or parasocial contexts represents a growing intersection between psychology and human-computer interaction. These areas may involve unique ethical, methodological, and clinical challenges that extend beyond the scope of our recommendations. Studies using LLMs in clinical settings, such as mental-health chatbots, should adhere to established clinical-research guidelines in addition to the best practices outlined here. Researchers are encouraged to consider guidelines such as CONsolidated Standards of Reporting Trials-Artificial Intelligence (X. Liu et al., 2020) for reporting randomized clinical trials, Standard Protocol Items: Recommendations for Interventional Trials-Artificial Intelligence (Rivera et al., 2020) for reporting trial protocols, Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis-Artificial Intelligence (Collins et al., 2024) for reporting diagnostic and prognostic model development, and Chatbot Assessment Reporting Tool (CHART) (CHART Collaborative, 2024), which is specifically designed for chatbot interventions. These frameworks provide essential standards for ensuring the safety, efficacy, and ethical use of AI-driven tools in clinical contexts, which may not be fully covered by the general guidelines for social-science research involving LLMs.

Given the rapid pace of development of the field, it is crucial to continually reassess and refine guidelines for LLM use in research as new applications and challenges emerge. For example, as LLM's visual and audio processing improves, future guidelines might include recommendations on how to handle multimodal inputs, such as combining text, images, and sounds in research tasks. As LLMs handle more information without significant performance loss, future guidelines should address challenges with coding massive data sets, such as large-scale qualitative data (e.g., years of diary entries from thousands of individuals).

Validating such large outputs may pose new challenges: How can the field ensure accuracy when human review, even on subsets of the data, is impractical or even impossible? Likewise, when LLMs become viable for large-scale complex simulations (e.g., very large multiagent frameworks), guides may distill and adopt methodologies from agent-based modeling (ABM) to better understand and validate LLM performance (e.g., analogous to classical ABM best practices in Hammond, 2015). We recommend that future updates to this primer incorporate specific considerations for these specialized areas to better support researchers and reviewers in maintaining the highest standards of research quality and ethical practice.

Conclusion

LLMs are powerful tools that can automate previously time-consuming or expensive tasks. They have the potential to expand the scope of social-science research by allowing for the analysis of larger qualitative data sets or for simulating human behavior. As LLMs become more accessible and affordable, we expect their use in research to grow, but as with any technology, LLMs must be used appropriately with a clear understanding of their limitations. Both researchers and reviewers will increasingly need to understand appropriate-use cases and best practices for LLMs.

In this article, we highlighted key processes for producing reliable, transparent, and valid social-science research using LLM-generated data. It is important to keep in mind that this technology is advancing rapidly and that best practices are likely to change in the future. Providing detailed information, such as computer code and thorough explanations of methods, is crucial to help future researchers understand and if necessary, revise research conducted with the current generation of LLMs. Challenging as it may be, the best authors and reviewers will need to keep abreast of the latest guidance for reviewing LLM-based social-science research. As with any new technology, LLMs open new horizons while generating new pitfalls, and in this primer, we aim to help researchers who study human behavior benefit from this technology while avoiding some of the methodological challenges they pose.

Appendix

In this appendix, we walk through a concrete example of applying our checklist (see Table 3). In the context of a hypothetical study, we provide detailed examples of how each of the items in our checklist might be addressed effectively.

Example walk-through: large-language-model coding of moral framing and stance in migration debates

Study context. Researchers want to investigate online debates regarding pro-immigration and anti-immigration stances, focusing on how each group uses moral language to express their positions. By analyzing social media posts using moral foundations theory (MFT; Graham et al., 2013), the researchers aims to understand the role of moral language in polarized debates and show how different moral foundations are employed to support various stances.

The researchers collected a corpus of social media posts in pro-immigration and anti-immigration discussions. They plan to use a large language model (LLM) to classify each post's stance (pro-immigration or anti-immigration) and the post's moral framing (e.g., individualizing vs. binding values).

1. **[Optional] Preregister the study**

1.1. **Are the methods, including models, parameters, and validation strategies, preregistered?**

- **Example:**
  - The researchers completed a preregistration on OSF, specifying that they will use a specific instance of ChatGPT (GPT-4o; gpt-4o-2024-08-06) to annotate social media posts.
  - They preregister the following:
    - A plan to code each post's stance (pro-immigration, anti-immigration, or neutral),
    - a plan to analyze moral framing using MFT categories (care/harm, fairness/cheating, loyalty/betrayal, authority/subversion, purity/degradation),
    - a plan for validating a specific subset of LLM outputs by comparing them with human annotations,
    - a note that they will finalize prompt design and model settings by iterating on a small subset of the human validation data to maximize accuracy and minimize bias before testing on the full validation data set.

1.2. **Does the preregistration allow for a full understanding of the intended experimental design, data-analysis plan, and how results will be interpreted?**

- **Example:**
  - **Hypothesis:** The researchers provide a clear hypothesis.

1. The researchers hypothesize that pro- and anti-immigration groups will frame their arguments using different moral foundations:
- anti-immigration using binding values
- pro-immigration using individualizing values
  - **Sampling for validation:** The researchers specify that they will randomly select 1,000 posts for human annotation to serve as validation data.

- **Annotation with LLM:** The researchers specify how the LLM will be used to annotate posts.
    1. Classify posts' stance as pro-immigration, anti-immigration, neutral.
    2. Classify posts' moral framing (care, fairness, loyalty, authority, purity).
- **LLM settings:** The researchers specify relevant model settings and provide example prompts and procedures.
    - They will submit one post at a time to the LLM.
    - They will set the temperature parameter to 1.
    - They provide examples of prompts they will use.
    - They are clear about the need to iteratively improve prompts and model settings once data analysis has started.
    - **Analysis:** The researchers specify exact analyses they will use to analyze LLM outputs.

1. They will analyze the relationship between moral foundations and immigration stances.
2. They specify statistical models, for example, logistic regression, in which moral foundations predict stance.
    - **Interpretation:** The researchers clearly state how the statistical model outputs will be interpreted and how they will support/fail to support the hypothesis.
        1. A positive coefficient for individualizing framing predicting pro-immigration stance will support the hypothesis.
        2. A positive coefficient for binding framing predicting anti-immigration stance will support the hypothesis.

1.3. How closely does the study follow the preregistered protocols, and are any changes justified with transparent reasoning?

- **Example:**
    - **Adherence:** The study followed the same sampling procedure, model usage, and moral-foundation definitions.
    - **Changes:** Researchers increased the human validation sample from 1,000 to 2,000 posts to address data imbalance (e.g., not enough pro-immigration posts in the original sample). They documented this change in the final report, explaining it was needed to improve statistical power. As expected (and specified in the preregistration), they also changed the prompt and temperature setting from the ones originally preregistered.

2. [Required] Check if the model is stable (i.e., does not change over time) and accessible (i.e., can be used for replication)

2.1. Is the model stable?

- **Example:**
    - No, the researchers used ChatGPT (GPT-4o), which receives updates and might produce different outputs at different times.

If the model is not stable:

1. **Provide a justification**
    - **Justification:** They require GPT-4o's more advanced reasoning capabilities and performance for the complex classifications. The model needs to determine both stance and moral foundations used in the context of the posts that may not explicitly mention immigration.
2. **Provide the model's exact name and query date**
    - **Example:** GPT-4o used between September 1, 2024, and September 30, 2024, model version "gpt-4o-2024-08-06."
3. **Provide limitations**
    - **Example:** Potential inconsistencies over time because of updates.
4. **[Optional] Validate with a stable model**
    - **Example:** The researchers might use a stable open-source LLM (e.g., LLaMA-3 locked checkpoint) on a subset of the data to see if results are roughly comparable.

2.2. Is the model accessible?

- **Example:**
    - Yes, GPT-4o is accessible via the OpenAI application programming interface (API), although it requires an API key and may incur costs.
    - The researchers decide that GPT-4o is the best option given performance but acknowledge that for replication, the model might not be accessible in the future (e.g., if being deprecated).

If the model is not accessible:

- **Justification of exclusion of accessible models:**
    - Not applicable here because GPT-4o is (currently) accessible. In case the model might be inaccessible in the future, the researchers may refer to their validation using a stable open-source model.

3. [Required] Provide all materials for replication

3.1. Codes: The researcher provides all codes and data in a GitHub repository.

3.2. Model parameters and settings

Example: The researcher provides all relevant parameters and model settings.

- **Model:** gpt-4o-2024-08-06
- Temperature: 0.7
- Other relevant parameters: set to default

3.3. Prompts: The researchers iterated their prompt design on a small subset of the validation data until they found a well-performing prompt and model settings. For example, they found a few-shot prompt design (includes examples of the respective classifications) that performed well. They provide the final prompts used in their study. Below we present their prompts for classifying posts based on MFT and classifying posts based on immigration stance:

MFT classifier:

- **System prompt:** "You are a helpful classifier."
- User prompt:

> You are a text classifier designed to analyze content based on Moral Foundations Theory (MFT). Your task is to determine if the following post expresses one or more of the following moral values:
>
> Care/harm: Concern for the well-being of others, preventing harm, or alleviating suffering.
>
> Fairness/cheating: Focus on justice, rights, and equality, or condemnation of unfair practices.
>
> Loyalty/betrayal: Emphasis on allegiance, loyalty to a group, or betrayal of one's community.
>
> Authority/subversion: Respect for traditions, rules, or social hierarchy, or rejection of such authority.
>
> Purity/degradation: Importance of purity, sanctity, or rejection of things considered impure or degrading.
>
> Here is the post: [POST TEXT]
>
> Return all expressed values comma separated and nothing else.

Stance classifier:

- **System prompt:** "You are a helpful classifier."
- User Prompt:

> You are a text classifier designed to analyze content related to immigration stances. Your task is to determine if the following post expresses support for either the pro-immigration or anti-immigration perspective or if no clear stance is detectable.
>
> Here is an example of an "Anti-Immigration" (the post supports restricting immigration or opposes immigration) post: [EXAMPLE].
>
> Here is an example of a "Pro-Immigration" (the post supports immigration rights or favors immigration) post: [EXAMPLE].
>
> Here is an example of an "Unclear/Neutral" (the post does not clearly express support for either perspective) post: [EXAMPLE].
>
> Here is the post: [POST TEXT]
>
> Return only the expressed stance (either "pro-immigration", "anti-immigration", "neutral").

3.4. Data for fine-tuning

- In this study, no fine-tuning was performed. If fine-tuning was used, they would provide the training data (e.g., a json or csv file) and complete code to execute the fine-tuning steps.

3.5 Any other study material (e.g., questionnaires, human validation data)

- **Example:** The researchers provide a folder with the human annotations used to determine accuracy, including the codebook that instructed the human annotators.
- If relevant, the researchers also provide annotators demographics, for example, to study annotator bias.

3.6. Ensure code runs without errors

- The repository contains all code files used to preprocess the social media data, classify it using the LLM, and run the statistical analysis in a streamlined manner (e.g., a single file that runs everything sequentially or instructions for how to run each individual code file).
- The repository contains instructions to create a local programming environment and install all necessary packages.
- The code has been tested in a local environment with the required libraries installed and produces the same results as reported (or within reasonable margins).

3.7. Ensure replicated results align with reported results by a reasonable margin
- The model is run multiple times on the same data set. Outputs may vary slightly, but aggregated results (e.g., frequency counts of moral foundations) remain within ±5%.

3.8. Discuss strategies to account for LLM randomness
1. **Aggregation type:**
   - They run the model five times on each post and take the majority vote for final classification.
2. **Justification for aggregation:**
   - This reduces the impact of single-run randomness and yields more consistent outputs.

4. [Required] Ensure the reported results and inferences are justified

4.1. Were the LLM outputs validated with human data or other justifiable data?
- **Example:**
  - Yes, they compared LLM outputs on 1,000 randomly selected posts with human annotators. They report an accuracy of 90%, F_1 score of 0.85, and Cohen's κ of .75, indicating substantial agreement with human annotators.

Is the achieved accuracy sufficient?

- Yes, it is comparable with typical human-human agreement in stance detection.

Is the achieved accuracy discussed (e.g., comparison with other methods)?

- The researchers include a brief discussion of the model's performance, for example, the high agreement with human coders and high accuracy compared with alternative methods (e.g., typical accuracy/F_1 scores in stance detection)

4.2. Does the research question require robustness to different prompt strategies and model settings?

If yes, are LLM outputs robust to prompting strategies and model settings?

- Robustness of prompting strategies and model settings is not critical because the main goal is to automate human coding. The researchers iterated on a small subset of the validation data—treated as a "training set" for prompt development, analogous to standard train/test splits in machine learning—to find the optimal prompt and model settings that maximize accuracy and minimize bias. This approach avoids overfitting prompting strategies to the full validation set, preventing reporting inflated LLM performance that might not generalize.

If not, are the applied prompt strategy and the model settings justified?

- The LLM is used to automate human coders that would be significantly more expensive and slower. Thus, the prime concern is achieving sufficient accuracy compared with human raters. A secondary concern is ensuring that the model's misclassifications are not biasing the interpretations of the hypothesis.
- **Example:**
  - The researchers used a small subset of the human data to find the most accurate prompt and model settings and then verified accuracy and bias on the full human-validation data.
    - The model achieved a high accuracy (90%) and F_1 score (0.85) and a Cohen's κ comparable with human annotators.
    - Using the human ground-truth data, the researchers check that the model's misclassifications are not more frequent for any classification class.
    - Using the human ground-truth data, the researchers check that the model's misclassifications are not more frequent for any classification class.
    - For example, they report the F_1 scores for each class (e.g., pro-immigration vs. anti-immigration) and moral framing (e.g., care, fairness, authority, loyalty, purity), finding that there are no significant discrepancies across classes.
      - If they found discrepancies, they need further testing to determine whether these influence the interpretation of their results. For example, if pro-immigration posts with binding framing are significantly more often misclassified than individualizing framing, this should caution them from interpreting that pro-immigration posts are more likely to be framed with individualizing framing because this could be an artifact of the LLM bias/errors.

5. **Evaluate data processing and error handling**

5.1. **Are the data processing and error handling reasonable?**

- **Example:**
  - The researchers removed duplicate posts and excluded those with fewer than 15 characters.
  - If the model outputs an invalid classification (e.g., not one of the specified categories), the classification is reattempted up to three times. If it still fails, the post is dropped.
  - They report the frequency of invalid classifications with the final prompt.

5.2. **Are the data processing and error handling biased toward the desired outcomes?**

- **Example:**
  - The outlier exclusion is based on post length and unparseable responses, not on stance or moral content.
  - They confirm no correlation between invalid classifications and stance or moral framing to avoid skewed inference (e.g., pro-immigration or anti-immigration content was more likely to be excluded, which could affect downstream statistical analysis).

Transparency

Action Editor: Aishwarya Rajesh
Editor: David A. Sbarra
Author Contributions

Suhaib Abdurahman: Conceptualization; Investigation; Methodology; Visualization; Writing – original draft.
Alireza Salkhordeh Ziabari: Conceptualization; Investigation; Methodology.
Alexander K. Moore: Conceptualization; Writing – original draft.
Daniel M. Bartels: Conceptualization.
Morteza Dehghani: Conceptualization; Supervision.

Declaration of Conflicting Interests
The author(s) declared that there were no conflicts of interest with respect to the authorship or the publication of this article.

Funding
This research was supported in part by the Defense Advanced Research Projects Agency, Influence Campaign Awareness and Sensemaking HR001121C0165 and the Air Force Office of Scientific Research A9550-23-1-0463.

ORCID iDs

Suhaib Abdurahman https://orcid.org/0000-0001-5615-0129
Alexander K. Moore https://orcid.org/0000-0003-4046-3117
Morteza Dehghani https://orcid.org/0000-0002-9478-4365

Acknowledgments

The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of the Defense Advanced Research Projects Agency or the U.S. government. The U.S. government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation therein.

Notes

1. Large language models (LLMs) broadly fall into two categories: autoencoding models (e.g., BERT) and autoregressive models (e.g., ChatGPT). Autoencoding models excel at contextual tasks, such as text classification. Autoregressive models, trained on sequential word prediction, are versatile for text generation, conversation, and even classification via natural-language instructions. Recent advances have led to the widespread adoption of autoregressive models, notably ChatGPT, in various domains. In this primer, the term "LLMs" specifically refers to autoregressive models.
2. Many "open" AI models provide publicly available weights, enabling local use and investigation of their internal workings, but not full training code or data—limiting replication of the model's creation. Because this primer focuses on running rather than training models, we treat these "open-weight" models as open-source. For a more nuanced discussion of LLM "openness," see Liesenfeld et al. (2023).
3. These platforms benchmark language models on various tasks, akin to standardized testing in psychology.
4. See all code, data, and instructions for replication in this GitHub repository: https://github.com/goytoom/llm_psychology_guide.

References

Abdurahman, S., Atari, M., Karimi-Malekabadi, F., Xue, M. J., Trager, J., Park, P. S., Golazizian, P., Omrani, A., & Dehghani, M. (2024). Perils and opportunities in using large language models in psychological research. *PNAS Nexus*, *3*(7), Article pgae245. https://doi.org/10.1093/pnasnexus/pgae245

Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., Avila, R., Babuschkin, I., Balaji, S., Balcom, V., Baltescu, P., Bao, H., Bavarian, M., Belgum, J., Bello, I., . . . Zoph, B. (2023). *Gpt-4 technical report*. arXiv. https://doi.org/10.48550/arXiv.2303.08774

Aher, G. V., Arriaga, R. I., & Kalai, A. T. (2023). Using large language models to simulate multiple humans and replicate human subject studies. In *ICML'23: Proceedings of the 40th International Conference on Machine Learning* (pp. 337–371). Association for Computing Machinery.

Almeida, G. F., Nunes, J. L., Engelmann, N., Wiegmann, A., & de Araújo, M. (2023). *Exploring the psychology of gpt-4's moral and legal reasoning*. arXiv. https://doi.org/10.48550/arXiv.2308.01264

Amin, M. M., Cambria, E., & Schuller, B. W. (2023). Will affective computing emerge from foundation models and general artificial intelligence? A first evaluation of ChatGPT. *IEEE Intelligent Systems*, *38*(2), 15–23.

Argyle, L. P., Busby, E. C., Fulda, N., Gubler, J. R., Rytting, C., & Wingate, D. (2023). Out of one, many: Using language models to simulate human samples. *Political Analysis*, *31*(3), 337–351.

Astekin, M., Hort, M., & Moonen, L. (2024). An exploratory study on how non-determinism in large language models affects log parsing. In *Proceedings of the ACM/IEEE 2nd International Workshop on Interpretability, Robustness, and Benchmarking in Neural Software Engineering* (pp. 13–18). ACM/IEEE.

Atari, M., Omrani, A., & Dehghani, M. (2023). *Contextualized construct representation: Leveraging psychometric scales to advance theory-driven text analysis*. PsyArXiv. https://doi.org/10.31234/osf.io/m93pd

Atari, M., Xue, M. J., Park, P. S., Blasi, D., & Henrich, J. (2023). *Which humans?* PsyArXiv. https://doi.org/10.31234/osf.io/5b26t

Bai, H., Voelkel, J., Eichstaedt, J., & Willer, R. (2023). *Artificial intelligence can persuade humans on political issues*. Research Square.

Bang, Y., Cahyawijaya, S., Lee, N., Dai, W., Su, D., Wilie, B., Lovenia, H., Ji, Z., Yu, T., Chung, W., Do, Q. V., Xu, Y., & Fung, P. (2023). A multitask, multilingual, multimodal evaluation of ChatGPT on reasoning, hallucination, and interactivity. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 675–718). Association for Computational Linguistics.

Beck, T., Schuff, H., Lauscher, A., & Gurevych, I. (2024). Sensitivity, performance, robustness: Deconstructing the effect of sociodemographic prompting. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 2589–2615). Association for Computational Linguistics.

Binz, M., & Schulz, E. (2023). Using cognitive psychology to understand GPT-3. *Proceedings of the National Academy of Sciences*, *120*(6), Article e2218523120. https://doi.org/10.1073/pnas.2218523120

Blyler, A. P., & Seligman, M. E. (2024). Personal narrative and stream of consciousness: An AI approach. *The Journal of Positive Psychology*, *19*(4), 592–598.

Boyd-Graber, J., Okazaki, N., & Rogers, A. (2023a). *ACL 2023 policy on AI writing assistance*. https://2023.aclweb.org/blog/ACL-2023-policy/

Boyd-Graber, J., Okazaki, N., & Rogers, A. (2023b). *ACL'23 peer review form*. https://2023.aclweb.org/blog/review-form/#reproducibility-and-ethics

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., . . . Amodei, D. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems*, *33*, 1877–1901.

CHART Collaborative. (2024). Protocol for the development of the Chatbot Assessment Reporting Tool (CHART) for clinical advice. *BMJ Open*, *14*(5), Article e081155. https://doi.org/10.1136/bmjopen-2023-081155

Chen, J., & Yang, D. (2023). Unlearn what you want to forget: Efficient unlearning for LLMs. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing* (pp. 12041–12052). Association for Computational Linguistics.

Chen, L., Zaharia, M., & Zou, J. (2023). Analyzing chatgpt's behavior shifts over time [Poster presentation]. R0-FoMo: Robustness of Few-Shot and Zero-Shot Learning in Large Foundation Models Workshop. https://neurips.cc/virtual/2023/76729

Chiang, C. H., & Lee, H. Y. (2023). Can large language models be an alternative to human evaluations? In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 15607–15631). Association for Computational Linguistics.

Chollet, F. (2019). *On the measure of intelligence*. arXiv. https://doi.org/10.48550/arXiv.1911.01547

Chollet, F., Knoop, M., Kamradt, G., & Landers, B. (2024). *Arc prize 2024: Technical report*. arXiv. https://doi.org/10.48550/arXiv.2412.04604

Coda-Forno, J., Witte, K., Jagadish, A. K., Binz, M., Akata, Z., & Schulz, E. (2023). *Inducing anxiety in large language models increases exploration and bias*. arXiv. https://doi.org/10.48550/arXiv.2304.1111

Collins, G. S., Moons, K. G., Dhiman, P., Riley, R. D., Beam, A. L., Van Calster, B., Ghassemi, M., Liu, X., Reitsma, J. B, van Smeden, M., Boulesteix, A.-L., Camaradou, J. C., Celi, L. A., Denaxas, S., Denniston, A. K., Ben, G., Golub, R. M., Harvey, H., Glocker, B., . . . Logullo, P. (2024). TRIPOD+ AI statement: Updated guidance for reporting clinical prediction models that use regression or machine learning methods. *The BMJ*, *385*, Article q902. https://doi.org/10.1136/bmj.q902

Coyne, S., & Sakaguchi, K. (2023). *An analysis of gpt-3's performance in grammatical error correction*. arXiv. https://doi.org/10.48550/arXiv.2303.14342

Dam, S. K., Hong, C. S., Qiao, Y., & Zhang, C. (2024). *A complete survey on LLM-based AI chatbots*. arXiv. https://doi.org/10.48550/arXiv.2406.16937

Davis, J., Van Bulck, L., Durieux, B. N., & Lindvall, C. (2024). The temperature feature of ChatGPT: Modifying creativity for clinical research. *JMIR Human Factors*, *11*(1), Article e53559. https://doi.org/10.2196/53559

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). *BERT: Pre-training of deep bidirectional transformers for language understanding*. arXiv. https://doi.org/10.48550/arXiv.1810.04805

Dijkstra, R., Genç, Z., Kayal, S., & Kamps, J. (2022). Reading comprehension quiz generation using generative pre-trained transformers. In *Proceedings of the Fourth International Workshop on Intelligent Textbooks 2022* (pp. 4–17). https://ceur-ws.org/Vol-3192/itb22_p1_full5439.pdf

Dillion, D., Tandon, N., Gu, Y., & Gray, K. (2023). Can AI language models replace human participants? *Trends in Cognitive Sciences*, *27*(7), 597–600. https://doi.org/10.1016/j.tics.2023.04.008

Doshi, A. R., & Hauser, O. (2023). Generative artificial intelligence enhances creativity. *SSRN*. https://doi.org/10.2139/ssrn.4535536

Durmus, E., Nyugen, K., Liao, T. I., Schiefer, N., Askell, A., Bakhtin, A., Chen, C., Hatfield-Dodds, Z., Hernandez, D., Joseph, N., Lovitt, L., McCandlish, S., Sikder, O., Tamkin, A., Thamkul, J., Kaplan, J., Clark, J., & Ganguli, D. (2023). *Towards measuring the representation of subjective global opinions in language models*. arXiv. https://doi.org/10.48550/arXiv.2306.16388

Eldan, R., & Russinovich, M. (2023). *Who's Harry Potter? Approximate unlearning in LLMs*. arXiv. https://doi.org/10.48550/arXiv.2310.02238

Ericsson, K. A., & Moxley, J. H. (2019). Thinking aloud during superior performance on tasks involving decision making.In M. Schulte-Mecklenbeck, A. Kuehberger, & J. G. Johnson (Eds.), *A handbook of process tracing methods* (pp. 286–301). Routledge.

Fix, E., & Hodges, J. L. (1951). *Discriminatory analysis. Nonparametric discrimination: Small sample performance* (Report A, 193008). University of California Berkeley.

Fujita, H., et al. (2022). Prompt sensitivity of language model for solving programming problems. In *New Trends in Intelligent Software Methodologies, Tools and Techniques: Proceedings of the 21st International Conference on New Trends in Intelligent Software Methodologies, Tools and Techniques (SoMeT_22)*. Frontiers in Artificial Intelligence and Applications.

Garten, J., Hoover, J., Johnson, K. M., Boghrati, R., Iskiwitch, C., & Dehghani, M. (2018). Dictionaries and distributions: Combining expert knowledge and large scale textual data content analysis: Distributed dictionary representation. *Behavior Research Methods*, *50*, 344–361.

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*(12), 86–92.

Gilardi, F., Alizadeh, M., & Kubli, M. (2023). ChatGPT outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences*, *120*(30), Article e2305016120. https://doi.org/10.1073/pnas.2305016120

Graham, J., Haidt, J., Koleva, S., Motyl, M., Iyer, R., Wojcik, S. P., & Ditto, P. H. (2013). Moral foundations theory: The pragmatic validity of moral pluralism. In P. Devine & A. Plant (Eds.), *Advances in experimental social psychology* (Vol. 47, pp. 55–130). Academic Press.

Guo, D., Yang, D., Zhang, H., Song, J., Zhang, R., Xu, R., Zhu, Q., Ma, S., Wang, P., Bi, X., Zhang, X., Yu, X., Wu, Y., Wu, Z. F., Gou, Z., Shao, Z., Li, Z., Gao, Z., Shao, Z., . . . He, Y. (2025). *Deepseek-r1: Incentivizing reasoning capability in LLMs via reinforcement learning*. arXiv. https://doi.org/10.48550/arXiv.2501.12948

Hammond, R. A. (2015). Considerations and best practices in agent-based modeling to inform policy. In R. Wallace, A. Geller, & V. A. Ogawa (Eds.), *Assessing the use of agent-based models for tobacco regulation* (Appendix A). National Academies Press.

He, J., Wallis, F., Gvirtz, A., & Rathje, S. (2024). Artificial intelligence chatbots mimic human collective behaviour. *British Journal of Psychology*. Advance online publication. https://doi.org/10.1111/bjop.12764

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). Beyond WEIRD: Towards a broad-based behavioral science. *Behavioral and Brain Sciences*, *33*(2–3), 111–135. https://doi.org/10.1017/S0140525X10000725

Hewitt, L., Ashokkumar, A., Ghezae, I., & Willer, R. (2024). *Predicting results of social science experiments using large language models* (Technical report, Working Paper). https://samim.io/dl/Predicting%20results%20of%20social%20science%20experiments%20using%20large%20language%20models.pdf

Horton, J. J. (2023). *Large language models as simulated economic agents: What can we learn from homo silicus?* (Technical report). National Bureau of Economic Research.

Hugging Face. (2024a). *Inference*. https://huggingface.co/docs/huggingface_hub/en/package_reference/inference_client#huggingface_hub.InferenceClient.chat_completion

Hugging Face. (2024b). *Open LLM leaderboard 2*. https://huggingface.co/spaces/HuggingFaceH4/open_llm_leaderboard

Hutchinson, B., Prabhakaran, V., Denton, E., Webster, K., Zhong, Y., & Denuyl, S. (2020). *Social biases in NLP models as barriers for persons with disabilities*. arXiv. https://doi.org/10.48550/arXiv.2005.0081

Jiang, A. Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D. S., Casas, D. d. l., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., Lavaud, L. R., Lachaux, M.-A., Stock, P., Le Scao, T., Lavril, T., Wang, T., Lacroix, T., & El Sayed, W. (2023). *Mistral 7b*. arXiv. https://doi.org/10.48550/arXiv.2310.06825

Ke, L., Tong, S., Chen, P., & Peng, K. (2024). *Exploring the frontiers of LLMs in psychological applications: A comprehensive review*. arXiv. https://doi.org/10.48550/arXiv.2401.01519

Kobak, D., González-Márquez, R., Horvát, E.-Á., & Lause, J. (2024). *Delving into ChatGPT usage in academic writing through excess vocabulary*. arXiv. https://doi.org/10.48550/arXiv.2406.07016

Kumar, P. (2024). Large language models (LLMs): Survey, technical frameworks, and future challenges. *Artificial Intelligence Review*, *57*, Article 260. https://doi.org/10.1007/s10462-024-10888-y

Le Scao, T., Fan, A., Akiki, C., Pavlick, E., Ilić, S., Hesslow, D., Castagné, R., Luccioni, A. S., Yvon, F., Gallé, M., Tow, J., Rush, A. M., Biderman, S., Webson, A., Ammanamanchi, P. S., Wang, T., Sagot, B., Muennighoff, N., Villanova del Moral, A., . . . Wolf, T. (2023). *Bloom: A 176b-parameter open-access multilingual language model*. arXiv. https://doi.org/10.48550/arXiv.2211.05100

Lee, S., Sim, W., Shin, D., Seo, W., Park, J., Lee, S., Hwang, S., Kim, S., & Kim, S. (2024). Reasoning abilities of large language models: In-depth analysis on the abstraction and reasoning corpus. In *ACM Transactions on Intelligent Systems and Technology*. Association for Computation Machinery. https://doi.org/10.1145/3712701

Liang, W., Zhang, Y., Wu, Z., Lepp, H., Ji, W., Zhao, X., Cao, H., Liu, S., He, S., Huang, Z., Yang, D., Potts, C., Manning, C. D., & Zou, J. Y. (2024). *Mapping the increasing use of LLMs in scientific papers*. arXiv. https://doi.org/10.48550/arXiv.2404.01268

Liesenfeld, A., Lopez, A., & Dingemanse, M. (2023). Opening up ChatGPT: Tracking openness, transparency, and accountability in instruction-tuned text generators. In *Proceedings* *of the 5th International Conference on Conversational User Interfaces* (pp. 1–6). Association for Computing Machinery.

Liu, A., Feng, B., Xue, B., Wang, B., Wu, B., Lu, C., Zhao, C., Deng, C., Zhang, C., Ruan, C., Dai, D., Guo, D., Yang, D., Chen, D., Ji, D., Li, E., Lin, F., Dai, F., Luo, F., . . . Pan, z. (2024). *Deepseek-v3 technical report*. arXiv. https://doi.org/10.48550/arXiv.2412.19437

Liu, H., Ning, R., Teng, Z., Liu, J., Zhou, Q., & Zhang, Y. (2023). *Evaluating the logical reasoning ability of ChatGPT and gpt-4*. arXiv. https://doi.org/10.48550/arXiv.2304.03439

Liu, X., Rivera, S. C., Moher, D., Calvert, M. J., & Denniston, A. K.; SPIRIT-AI and CONSORT-AI Working Group. (2020). Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: The CONSORT-AI extension. *The Lancet Digital Health*, *2*(10), e537–e548.

Liu, Z., Dou, G., Tan, Z., Tian, Y., & Jiang, M. (2024). *Towards safer large language models through machine unlearning*. arXiv. https://doi.org/10.48550/arXiv.2402.1005

LMSYS. (2024). *Chatbot Arena Leaderboard updates (Week 4)*. https://chat.lmsys.org/?leaderboard

Loya, M., Sinha, D., & Futrell, R. (2023). Exploring the sensitivity of LLMs' decision-making capabilities: Insights from prompt variations and hyperparameters. In H. Bouamor, J. Pino, & K. Bali (Eds.), *Findings of the Association for Computational Linguistics: EMNLP 2023* (pp. 3711–3716). Association for Computational Linguistics.

Lu, Y., Bartolo, M., Moore, A., Riedel, S., & Stenetorp, P. (2022). Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 8086–8098). Association for Computational Linguistics.

Maini, P., Feng, Z., Schwarzschild, A., Lipton, Z. C., & Kolter, J. Z. (2024). *TOFU: A task of fictitious unlearning for LLMs*. arXiv. https://doi.org/10.48550/arXiv.2401.06121

Matter, D., Schirmer, M., Grinberg, N., & Pfeffer, J. (2024). Investigating the increase of violent speech in Incel communities with human-guided GPT-4 prompt iteration. *Frontiers in Social Psychology*, *2*, Article 1383152. https://doi.org/10.3389/frsps.2024.1383152

Messeri, L., & Crockett, M. J. (2024). Artificial intelligence and illusions of understanding in scientific research. *Nature*, *627*(8002), 49–58.

Minaee, S., Mikolov, T., Nikzad, N., Chenaghlu, M., Socher, R., Amatriain, X., & Gao, J. (2024). *Large language models: A survey*. arXiv. https://doi.org/10.48550/arXiv.2402.06196

Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I. D., & Gebru, T. (2019). Model cards for model reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency* (pp. 220–229). Association for Computing Machinery.

Monniaux, D. (2008). The pitfalls of verifying floating-point computations. *ACM Transactions on Programming Languages and Systems (TOPLAS)*, *30*(3), 1–41.

Naismith, B., Mulcaire, P., & Burstein, J. (2023). Automated evaluation of written discourse coherence using GPT-4. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)* (pp. 394–403). Association for Computational Linguistics.

OpenAI. (2023). *Advanced usage - OpenAI API*. https://platform.openai.com/docs/advanced-usage/reproducible-outputs

OpenAI. (2024). *Introducing OpenAI o1-preview*. https://openai.com/index/introducing-openai-o1-preview/

Ouyang, S., Zhang, J. M., Harman, M., & Wang, M. (2023). *LLM is like a box of chocolates: The non-determinism of ChatGPT in code generation*. arXiv. https://doi.org/10.48550/arXiv.2308.02828

Park, J. S., O'Brien, J., Cai, C. J., Morris, M. R., Liang, P., & Bernstein, M. S. (2023). Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology* (pp. 1–22). Association for Computing Machinery.

Park, J. S., Popowski, L., Cai, C., Morris, M. R., Liang, P., & Bernstein, M. S. (2022). Social simulacra: Creating populated prototypes for social computing systems. In *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology* (pp. 1–18). Association for Computing Machinery.

Park, P. S., Schoenegger, P., & Zhu, C. (2024). Diminished diversity-of-thought in a standard large language model. *Behavior Research Methods*, *56*, 5754–5770. https://doi.org/10.3758/s13428-023-02307-x

Pawelczyk, M., Neel, S., & Lakkaraju, H. (2023). *In-context unlearning: Language models as few shot unlearners*. arXiv. https://doi.org/10.48550/arXiv.2310.07579

Pennebaker, J. W., Booth, R. J., & Francis, M. E. (2007). *Linguistic inquiry and word count: LIWC* [Computer software]. LIWC.

Puigcerver, J., Ruiz, C. R., Mustafa, B., & Houlsby, N. (2023). *From sparse to soft mixtures of experts* [Conference session]. The Twelfth International Conference on Learning Representations, Vienna, Austria.

Rathje, S., Mirea, D. M., Sucholutsky, I., Marjieh, R., Robertson, C. E., & Van Bavel, J. J. (2024). GPT is an effective tool for multilingual psychological text analysis. *Proceedings of the National Academy of Sciences*, *121*(34), Article e2308950121. https://doi.org/10.1073/pnas.2308950121

Rivera, S. C., Liu, X., Chan, A. W., Denniston, A. K., & Calvert, M. J.; The SPIRIT-AI and CONSORT-AI Working Group. (2020). Guidelines for clinical trial protocols for interventions involving artificial intelligence: The SPIRIT-AI extension. *The Lancet Digital Health*, *2*(10), e549–e560.

Santurkar, S., Durmus, E., Ladhak, F., Lee, C., Liang, P., & Hashimoto, T. (2023). *Whose opinions do language models reflect?* [Conference session]. International Conference on Machine Learning, Honolulu, Hawaii.

Sclar, M., Choi, Y., Tsvetkov, Y., & Suhr, A. (2024). *Quantifying language models' sensitivity to spurious features in prompt design or: How I learned to start worrying about prompt formatting* [Conference session]. The Twelfth International Conference on Learning Representations, Vienna, Austria.

Shanahan, M. (2024). Talking about large language models. *Communications of the ACM*, *67*(2), 68–79.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, *22*(11), 1359–1366.

Suri, G., Slater, L. R., Ziaee, A., & Nguyen, M. (2024). Do large language models show decision heuristics similar to humans? A case study using GPT-3.5. *Journal of Experimental Psychology: General*, *153*(4), 1066–1075. https://doi.org/10.1037/xge0001547

Tabone, W., & de Winter, J. (2023). Using ChatGPT for human–computer interaction research: A primer. *Royal Society Open Science*, *10*(9), Article 231053. https://doi.org/10.1098/rsos.231053

Thaker, P., Maurya, Y., & Smith, V. (2024). *Guardrail baselines for unlearning in LLMs*. arXiv. https://doi.org/10.48550/arXiv.2403.03329

Törnberg, P. (2023). *ChatGPT-4 outperforms experts and crowd workers in annotating political twitter messages with zero-shot learning*. arXiv. https://doi.org/10.48550/arXiv.2304.06588

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., & Rodriguez, A. (2023). *LLaMA: Open and efficient foundation language models*. arXiv. https://doi.org/10.48550/arXiv.2302.13971

Wang, A., Morgenstern, J., & Dickerson, J. P. (2024). *Large language models should not replace human participants because they can misportray and flatten identity groups*. ArXiv. https://doi.org/10.48550/arXiv.2402.01908

Wang, Y., Yao, Q., Kwok, J. T., & Ni, L. M. (2020). Generalizing from a few examples: A survey on few-shot learning. *ACM Computing Surveys (CSUR)*, *53*(3), 1–34. https://doi.org/10.1145/3386252

Wang, Z. P., Bhandary, P., Wang, Y., & Moore, J. H. (2024). Using GPT-4 to write a scientific review article: a pilot evaluation study. *BioData Mining*, *17*(1), Article 16. https://doi.org/10.1186/s13040-024-00371-3

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q. V., & Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, *35*, 24824–24837.

Wu, S., Fei, H., Qu, L., Ji, W., & Chua, T. S. (2024). Next-GPT: Any-to-any multimodal LLM. In *ICML'24: Proceedings of the 41st International Conference on Machine Learning* (pp. 53366–53397). Association for Computing Machinery.

Xiao, G., Liu, J., Zheng, Z., & Sui, Y. (2021). Nondeterministic impact of CPU multithreading on training deep learning systems. In *2021 IEEE 32nd International Symposium on Software Reliability Engineering (ISSRE)* (pp. 557–568). IEEE.

Xu, Y., Li, W., Vaezipoor, P., Sanner, S., & Khalil, E. B. (2023). LLMs and the abstraction and reasoning corpus: Successes, failures, and the importance of object-based representations. *Transactions on Machine Learning Research*. arXiv. https://doi.org/10.48550/arXiv.2305.18354

Yin, S., Fu, C., Zhao, S., Li, K., Sun, X., Xu, T., & Chen, E. (2023). *A survey on multimodal large language models*. arXiv. https://doi.org/10.48550/arXiv.2306.13549

Zhang, D., Finckenberg-Broman, P., Hoang, T., Pan, S., Xing, Z., Staples, M., & Xu, X. (2023). *Right to be forgotten in the era of large language models: Implications, challenges, and solutions*. arXiv . https://doi.org/10.48550/arXiv.2307.03941

Zhao, Y., Zhang, R., Li, W., Huang, D., Guo, J., Peng, S., Hao, Y., Wen, Y., Hu, X., Du, Z., Guo, Q., & Li, L. (2024). *Assessing and understanding creativity in large language models*. arXiv. https://doi.org/10.48550/arXiv.2401.12491